CLASSIFICATION OF AFFECT USING NOVEL VOICE AND VISUAL FEATURES

A Dissertation
Presented to
The Academic Faculty

By

Jonathan Chongkang Kim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2014

CLASSIFICATION OF AFFECT USING NOVEL VOICE AND VISUAL FEATURES

Approved by:

Dr. Mark A. Clements, Advisor
Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. David V. Anderson
Professor, School of School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Elliot Moore II
Associate Professor, School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Agata Rozga
Research Scientist, School of Interactive Computing
Georgia Institute of Technology

Dr. Irfan Essa
Professor, School of Interactive Computing
Georgia Institute of Technology

Date Approved: October 2, 2014

To my wife and my daughter

ACKNOWLEDGMENTS

I simply could not complete this dissertation without help and support from my advisor Prof. Mark Clements. I still remember the day when I visited Prof. Clements' office with my résumé for a research assistantship position in his lab. He did not even take a look at my résumé which only included a few lines about my signal processing experiences I had from taking his course a year before. I still question what made him trust me and my ability for the Ph.D. completion. I thank him for trusting me and allowing me the freedom to explore the vast area of signal processing.

The quality of this dissertation has been "statistically significantly" improved by the insights and comments from the committee members. Prof. Moore was the first one who introduced me the field of affective computing. I will always remember the night we had to finish up a conference paper which was due the very next morning. When I was working on an audio signal modification project, Prof. Anderson did not hesitate to share with me his MATLAB codes he wrote while he was a Ph.D. student. Prof. Essa has always encouraged me and provided deep insights during my presentations for the Expedition project which was my main source of funding for last four years. I thank Dr. Rozga, the mother of the Child Study Lab, for reviewing my conference and journal papers and this dissertation with a keen eye.

In addition to my advisor and the committee members, there are so many other professors I must give thanks to. I thank Profs. Rehg and Abowd for helping me see a problem with different angles. Their comments in the Expedition meetings were truly valuable. I thank Prof. McClellan for giving me the teaching assistantship opportunity for ECE 2025. He also helped me build a strong foundation of signal processing. I thank Prof. Bruce Walker for giving me an opportunity to be a part of his aquarium project. I truly enjoyed using my music skills in a science project.

I can never forget the many memories with my friends at the Center for Signal and

Image Processing (CSIP). There are so many people responsible for these memories, and I cannot name them all here. I will put a few of them in a chronological order of when I met them: Yeongseon Lee, Jinwoo Kang, Kaustubh Kalgaonkar, Brett Matthews, Soohyun Bae, Byungki Byun, Antonio Moreno, Ted Wada, Ilseo Kim, Sunghwan Shin, Seok-Chul Kwon, Haejoon Jung, Hrishikesh Rao, Asif Ali, Teresa Sanders, and Hyunwoo Cho. Our pain and suffering at the CSIP will finally pay off (or have already paid off for those who got out of here before me.)

I owe so much to my family. I thank my parents and parents-in-law for their financial and moral support. I thank Dr. Joseph Kim for being a good brother and a math tutor. Without his after-school math tutoring sessions, I would not have been able to come to Georgia Tech from high school. I thank my sisters-in-law for their encouragement and for making me laugh with their silly jokes. Of course, I cannot forget to mention my wife, Jiyoung. I thank her for being patient through this long process and for constantly reminding me that our budget is tight. Lastly but most importantly, I thank God the Almighty for who He is and what He is doing in and through my life.

TABLE OF CONTENTS

ACKNOWLEDGMENTS iv

LIST OF TABLES ix

LIST OF FIGURES xii

SUMMARY xiv

CHAPTER 1 INTRODUCTION 1
1.1 Theories of Emotion 2
1.2 Databases of Emotional Speech 3
1.3 Baseline Features 4
1.4 Emotion Classifiers 5
1.5 Summary 8
1.6 Organization 9

CHAPTER 2 FORMANT-BASED FEATURES FOR EMOTION CLASSIFICATION 11
2.1 Formant-based Feature Extraction using Linear Predictive Coding 12
2.1.1 Feature Selection 13
2.2 Evaluation on Formant-based Features Extracted using LPC 17
2.2.1 Classification on the Development Set 19
2.3 Formant Frequency Tracking using Gaussian Mixtures 20
2.3.1 Parameter Initialization by Expectation-Maximization 21
2.3.2 Parameter Estimation by MAP Adaptation 24
2.3.3 Formant Frequency Tracking Results 25
2.4 Evaluation of Formant-Based Features Extracted Using GMM 29
2.4.1 Geneva Multimodal Emotion Portrayals Corpus 29
2.4.2 Classification Results 31
2.5 Conclusion 35

CHAPTER 3 SPECTRAL FEATURE EXTRACTION USING MULTI-RESOLUTION SINUSOIDAL TRANSFORM CODING 37
3.1 Application of Spectral Features to Affect Classification 38
3.1.1 Feature Extraction using MRSTC 38
3.1.2 Classification Results 43
3.2 Conclusion 46

CHAPTER 4 EMOTIONAL SPEECH ANALYSIS AT VARIOUS TEMPORAL LENGTHS 49
4.1 Classifier Fusion with Binary Matrices 50
4.1.1 Spectral Clustering using a Similarity Matrix 53
4.1.2 Classification Score 54

4.2 Experiments and Results . . . 55
4.2.1 Experiment I: Classification without Fusion . . . 55
4.2.2 Experiment II: Classification with Fusion . . . 57
4.2.3 Experiment III: Fusion of Fusion . . . 58
4.3 Conclusion . . . 59

CHAPTER 5 MULTIMODAL-MULTITEMPORAL EMOTION CLASSIFICATION . . . 60
5.1 IEMOCAP Database . . . 61
5.2 Feature Extraction . . . 62
5.2.1 Speech Feature Extraction . . . 62
5.2.2 MoCap Feature Extraction . . . 64
5.3 Unimodal-Unitemporal Classifiers . . . 66
5.3.1 SVMs for Imbalanced Dataset . . . 68
5.3.2 Feature Analysis . . . 69
5.3.3 Unimodal-Unitemporal Classification Results . . . 74
5.4 Multimodal-Multitemporal Classifier Fusion . . . 75
5.5 Fusion Results . . . 76
5.5.1 Diversity Measures . . . 76
5.5.2 Unimodal-Multitemporal Fusion Results . . . 78
5.5.3 Multimodal-Multitemporal Fusion Results . . . 79
5.5.4 Comparison to Context-Sensitive Classifiers . . . 84
5.6 Non-linguistic Vocalizations for Emotion Recognition . . . 85
5.6.1 Paralinguistic in the IEMOCAP Database . . . 86
5.6.2 Paralinguistic Classification . . . 88
5.6.3 Paralinguistic for Emotion Classification . . . 90
5.7 Conclusion and Discussion . . . 92

CHAPTER 6 SOCIAL ENGAGEMENT CLASSIFICATION IN DYADIC PLAYS . . . 95
6.1 Multimodal Dyadic Behavior Dataset . . . 96
6.2 Automatic Voice Annotation . . . 99
6.2.1 Voice Activity Detection . . . 100
6.2.2 Cross-talk Detection . . . 101
6.2.3 Paralinguistic Detection . . . 103
6.3 Feature Extraction . . . 104
6.3.1 Local Binary Features . . . 105
6.3.2 Stage-level Features . . . 106
6.4 Classifier for Engagement Classification . . . 106
6.5 Results . . . 108
6.6 Conclusion . . . 112

CHAPTER 7 SUMMARY AND CONCLUSION . . . 114

APPENDIX A APPLICATION OF SPECTRAL FEATURES IN SPEECH INTELLIGIBILITY ESTIMATION 119
A.1 NKI CCRT Speech Corpus 119
A.2 Intelligibility Score Prediction 120
A.2.1 Binary intelligibility classification 121
A.2.2 Regression Analysis 126
A.3 Conclusion 129

REFERENCES 130

LIST OF TABLES

Table 1 Recently proposed emotion classifiers. 6

Table 2 Formant-related low-level descriptors (LLDs). 14

Table 3 List of statistical and regression measures for formant LLDs. 14

Table 4 Formant-based features selected using MaxARC, MaxRel and mRMR in the training set. 18

Table 5 Average recall rate on the training set using the three selection criteria and a 10-fold cross-validation technique. 19

Table 6 Classification results on the development set. ARC represents average recall rate while ACC stands for weighted accuracy. Base+Form is the combination of baseline and formant-based features. 19

Table 7 RMSE of the proposed method (GMM+MAP), GMM+cep, PRAAT, and MB for the first three formants. 27

Table 8 Unweighted average accuracy over 10 folds using the formant-based features analyzed with 400-ms windows using the proposed method (GMM+MAP) and PRAAT in the GEMEP database. 31

Table 9 Unweighted average accuracy over 10 folds using SVMs (C=1) with the combination of formant-based and baseline features analyzed with 400-ms windows. 32

Table 10 Confusion matrices of classifying two levels of activation and valence, using the combination of baseline and formant-based features analyzed with 400-ms windows. The classification was done at the utterance level using SVMs (C=1). 34

Table 11 Confusion matrix of classifying 12 categories of emotion, using the combination of baseline and formant-based features analyzed with 400-ms windows. The classification was done at the utterance level using SVMs (C=1). 34

Table 12 List of statistical measures for MRSTC feature extraction. 42

Table 13 Unweighted average accuracy over 10 folds using SVMs (C=1) with the combinations of baseline, formant, and spectral feature sets extracted at the 400-ms and utterance levels. 44

Table 14 Confusion matrices of classifying two levels of activation and valence, using the combination of baseline, formant-based, and MRSTC features analyzed with 400-ms and utterance levels. The classification was done at the utterance level using SVMs (C=1). 47

Table 15 Classification and detection results on a disjoint set using a Bayesian classifier (GMM) and the proposed method (BinF) at two temporal analysis lengths before fusion. 56

Table 16 Classification and detection results in unweighted accuracy (UWA) using the proposed method with fusion and percentage points of improvement by fusion. 58

Table 17 Classification results in unweighted accuracy (UWA) after grand fusion and percentage points of improvement by grand fusion. 58

Table 18 List of statistical and regression measures applied the formant-related LLDs and the MoCap markers. 64

Table 19 Six groups of acoustic-prosodic LLDs with the numbers of extracted features. 65

Table 20 The distribution of the IEMOCAP database with the 3-level scale in the three emotional dimensions. 67

Table 21 The average UWAs $\pm$ σ of the seven individual unimodal-unitemporal classifiers for utterance-level and chunk-level classification. 75

Table 22 The average Q-values of paired classifiers over the three emotional dimensions. 78

Table 23 The UWAs of the unimodal-Multitemporal fusion and the best unitemporal classifier in each modality. 79

Table 24 The UWAs of the multimodal-multitemporal fusion using the proposed method (binF) and the best unimodal-unitemporal classifier in each dimension. 83

Table 25 Confusion matrices of the proposed Multimodal-multitemporal fusion method in three emotional dimensions. 84

Table 26 Audiovisual emotion classification unweighted accuracy at the utterance level using the proposed multimodal-multitemporal approach and the context-sensitive method proposed by [1]. 85

Table 27 The distribution of laughter and sighs in the IEMOCAP database in three emotional dimensions. 87

Table 28 Confusion matrices of the paralinguistic classifiers using speech and MoCap features. 88

Table 29 Emotion classification results using paralinguistic cues for chunk (400 ms) and utterance-level classification. 91

Table 30 Engagement score distribution in the 5 activities of 75 Rapid-ABC sessions. 98

Table 31 Confusion matrix of automatic segmentation in time (sec). 101

Table 32 Confusion matrix of cross-talk detection on segment level. 102

Table 33 Confusion matrices of the laughter and crying/fussing detectors. 104

Table 34 Eight local binary features from the annotations. 105

Table 35 Stage-level features from the annotations. 107

Table 36 Binary classification results in unweighted accuracy for four dyadic activities in 75 Rapid-ABC sessions using leave-one-session-out cross-validation with local and stage-level feature sets. 109

Table 37 Temporal fusion results in unweighted accuracy for four dyadic activities in 75 Rapid-ABC sessions using leave-one-session-out cross-validation. . 111

Table 38 The first 10 features in the optimal subsets of MRSTC and combined features. 123

Table 39 10-fold cross validation results on the training set using the optimal feature subsets. 124

Table 40 Classification results on the development set. 124

Table 41 Binary classification results on the test set. 126

Table 42 Confusion matrix of the SVM trained with the baseline and MRSTC features. 126

Table 43 Confusion matrices of the PLS regression predictions mapped into the 3 intelligibility levels. Chance would be 33.3. 129

LIST OF FIGURES

Figure 1 10-pole LPC response of the vowel /o/. 13

Figure 2 Error rate vs. number of features for MaxRel, mRMR, and MaxArc using a 10-fold cross-validation method on the training set. 17

Figure 3 Average recall (unweighted accuracy) vs. number of features to indicate the formant-based features selected by the feature selection criteria in the training set using a 10-fold cross-validation method. 18

Figure 4 System overview of the proposed method with a sequence of M analysis frames. 22

Figure 5 Spectrograms of two utterances with formants estimated by the proposed method (filled circles) and the hand-labeled ground truth (empty circles). 26

Figure 6 RMSE at various SNRs with (a) white noise and (b) babble for the sonorant-phonetic classes. 28

Figure 7 Harmonic peaks selected by SEEVOC peak-picking routine. 39

Figure 8 Discrete wavelet tiling for four-band MRSTC [2]. 40

Figure 9 A quadrature mirror filter analysis bank arranged for four-band MRSTC [2]. 40

Figure 10 STFT of normal, normal+jitter, and normal+noise signals in a low frequency band. 41

Figure 11 $R_{FF}(f_{lag})$ of normal, normal+jitter, and normal+noise signals. 42

Figure 12 The proposed feature extraction method overview. 43

Figure 13 Unweighted accuracy using SVMs (C=1) with baseline, formant, and spectral feature sets extracted at the 400-ms and utterance levels for (a) activation and (b) valence dimensions. 46

Figure 14 A graphical example of $\mathbf{A}_m$ matrix with 2 analysis lengths and 4 Gaussian clusters. 51

Figure 15 An example of the spectral clustering with four dimension-weighted binary vectors. 54

Figure 16 An example of the fused binary matrix $\mathbf{A}_m$ for an emotional state, interest. 57

Figure 17 MoCap markers are grouped into 6 facial regions. (The figure is adapted from [3].) 66

Figure 18 The proportions and the total number of the selected (a) speech features analyzed at the 400-ms, 800-ms, and utterance levels, and (b) MoCap features at the 50-ms, 400-ms, 800-ms, and utterance levels in the three emotional dimensions. 71

Figure 19 A graphical example of $\mathbf{A}_m$ matrix with 3 classifiers and 3 classes. 76

Figure 20 The sequential multimodal-multitemporal fusion results, where the order of fusion is from the classifier with the highest UWA (leftmost) to the classifier with the lowest UWA (rightmost). 81

Figure 21 The posterior probabilities of affective dimensions given laughter and sighs detected by speech-based and MoCap-based paralinguistic classifiers. The results using the ground truth labels are also included. Chance would be 33.3. 89

Figure 22 Screenshot of annotated Rapid-ABC session using Elan with continuous annotations of the subjects' behaviors. 98

Figure 23 Graphical user interface (GUI) developed using Matlab for voice segmentation and cross-talk detection. 103

Figure 24 Voice and cross-talk detection results with timestamps, which can be exported to a CSV file. 104

Figure 25 An example of local binary feature matrix from the annotations. 106

Figure 26 The posterior probabilities of engagement given each discrete behavioral event. The ground truth labels of the events are used. Chance would be 50.0. 108

Figure 27 Multitemporal fusion using the local binary features and the classification outputs of stage-level classifiers. 111

Figure 28 Feature selection results in unweighted accuracy with the number of features in each subset. 122

Figure 29 (top) Histogram of intelligibility scores in the development set. (bottom) Unweighted accuracy over intelligibility scores. 125

Figure 30 MSE of the (left) PCR and (right) PLS estimators with the number of predictor components. The shaded area represents the 95% confidence intervals. 127

Figure 31 Ground-truth intelligibility scores against the predicted scores of PLS. The Pearson's r value is 0.75. 128

SUMMARY

Emotion adds an important element to the discussion of how information is conveyed and processed by humans; indeed, it plays an important role in the contextual understanding of messages. This research is centered on investigating relevant features for affect classification, along with modeling the multimodal and multitemporal nature of emotion.

The use of formant-based features for affect classification is explored. Since linear predictive coding (LPC) based formant estimators often encounter problems with modeling speech elements, such as nasalized phonemes and give inconsistent results for bandwidth estimation, a robust formant-tracking algorithm was introduced to better model the formant and spectral properties of speech. The algorithm utilizes Gaussian mixtures to estimate spectral parameters and refines the estimates using maximum a posteriori (MAP) adaptation. When the method was used for features extraction applied to emotion classification, the results indicate that an improved formant-tracking method will also provide improved emotion classification accuracy.

Spectral features contain rich information about expressivity and emotion. However, most of the recent work in affective computing has not progressed beyond analyzing the mel-frequency cepstral coefficients (MFCC's) and their derivatives. A novel method for characterizing spectral peaks was introduced. The method uses a multi-resolution sinusoidal transform coding (MRSTC). Because of MRSTC's high precision in representing spectral features, including preservation of high frequency content not present in the MFCC's, additional resolving power was demonstrated.

Facial expressions were analyzed using 53 motion capture (MoCap) markers. Statistical and regression measures of these markers were used for emotion classification along the voice features. Since different modalities use different sampling frequencies and analysis window lengths, a novel classifier fusion algorithm was introduced. This algorithm is intended to integrate classifiers trained at various analysis lengths, as well as those obtained

from other modalities. Classification accuracy was statistically significantly improved using a multimodal-multitemporal approach with the introduced classifier fusion method.

A practical application of the techniques for emotion classification was explored using social dyadic plays between a child and an adult. The Multimodal Dyadic Behavior (MMDB) dataset was used to automatically predict young children's levels of engagement using linguistic and non-linguistic vocal cues along with visual cues, such as direction of a child's gaze or a child's gestures.

Although this and similar research is limited by inconsistent subjective boundaries, and differing theoretical definitions of emotion, a significant step toward successful emotion classification has been demonstrated; key to the progress has been via novel voice and visual features and a newly developed multimodal-multitemporal approach.

CHAPTER 1

INTRODUCTION

Communication is defined as the mutual exchange of information or ideas, including verbal communication and various forms of non-verbal communication such as vision, gesture, and expression. Since verbal communication is considered the fastest and most efficient method of communication, there has been much research on speech recognition since the late 1950s [4].

Recently, efforts in the development of speech recognition systems have come to fruition with an overflow of applications in our daily lives. However, despite the great success in speech recognition, we are still far from achieving natural interaction between man and machine, given that machines do not take into account the emotional state of speakers [4, 5]. Without changing the current human computing paradigm, where the transmission of explicit messages is emphasized while ignoring implicit information, interaction between humans and machines would not go beyond commands and responses [6]. In many applications, automatic transcription of speech requires descriptive information beyond what has merely been said. In human behavior analysis, such as measuring the engagement level of students in class and screening for autism spectrum disorder, automatic emotion classifiers would be useful.

In the realm of human communication, affect or emotion adds an important element to the discussion of how information is conveyed and processed by humans; indeed, it plays an important role in the contextual understanding of messages [7, 8, 9]. Humans express their feelings through different forms of communication, primarily speech, gestures, writing, and behavior [10, 1]. Speech, the most basic form of communication, transmits affective information through explicit linguistic messages, as well as through implicit paralinguistic features [5]. Automatic emotion classification has recently received attention due to its numerous areas of application. Example applications include psychiatric diagnosis, customer

relationship management, and children's behavior analysis [11, 7, 12].

Affect classification is a very challenging problem because emotions are constructs with fuzzy boundaries, and the acquisition of data with reliable labels is not easy. However, recent studies show that affect classifiers can be improved by including relevant features [7, 13], fusing multimodal features [14], using multi-layer classifiers [15], and employing physiological sensors [16, 17]. The emphasis of this dissertation will be on improving emotion classification by using novel and improved acoustic features and combining features from other modalities. Signal processing will be an important aspect of this research.

1.1 Theories of Emotion

Although affective computing is an emerging discipline, studies in the area of emotion go back to the 19th century, when Charles Darwin proposed his theory of emotions as expressions [5, 18]. According to Darwin, emotional expressions were initially associated with essential actions (e.g., a disgusted face was initially associated with rejecting an offensive object) [5, 18]. However, his theory fails to explain a number of emotional behaviors and expressions.

Darwin's perspective on emotions has been expanded by William James, who viewed emotions as embodiments or combinations of expression and physiological changes [5, 19]. The patterns of physiological changes associated with each emotion must be analyzed to truly understand one's emotional state.

In cognitive approaches, the person's expectations and goals in relation to the situation must be known to predict how a person will react to a situation [5, 20]. For a person to experience an emotion, an object or event must be appraised as directly affecting the person. Therefore, to understand his/her emotional state, a person's experience of a situation must be known.

James Averill viewed emotions as a social construct [21]. A social level of analysis is necessary to truly understand the nature of emotion. According to Averill, it is not

necessary for anger to be associated with aggression. Instead, expressions of anger serve as an agent to achieve certain social goals. Moreover, expressions of emotion are cultural. For instance, other cultures might have labels that literally cannot be translated into English [5, 21]. According to Pamela Cole's studies, expressing negative emotion may be viewed by Americans as appropriately assertive, but by the Napalese as woefully inappropriate. Cole's studies indicate that perception, as well as the expression of emotion, is social and cultural [22, 23].

In affective neuroscience, researchers propose methods to understand emotional processes and their neural correlates [5]. Neuroscientists often use functional magnetic resonance imaging (fMRI) techniques to provide new evidence into emotional phenomena [24]. In James Russell's view, emotion, as psychological construct, is a heterogeneous cluster of loosely related events, patterns, and dispositions [5, 25].

1.2 Databases of Emotional Speech

The most challenging task in constructing emotional speech databases is probably to elicit authentic affective expressions. Authentic affective expressions are very rare, and difficult to collect in studio settings. On the other hand, data collection in real-life settings is expensive and time-consuming. For this reason, most of the current emotional speech databases utilize actors for their data collection.

One of the main shortcomings of acted data is that they are often exaggerated and not sufficiently spontaneous. Despite the lack of spontaneity, acted data offer numerous advantages. Emotions underlying acted data can be defined precisely [26], and they offer the possibility of recording variable expressions by the same individuals. With "natural" expressions, only a few emotional reactions can be recorded for a given individual. To collect more spontaneous data, the authors of such data often employed the Stanislavski or Method acting technique. Hence, expressions emerge more spontaneously from the emotional state the actors have tried to produce, and "believable" emotions can be drawn

[26, 27, 3].

El Ayadi and Zhihong Zeng conducted a survey on 34 emotional speech databases [4, 6]. Thirteen of them consist of natural (spontaneous) speech data, and three out of the thirteen natural databases contain data collected at call centers. The sampling rate collected at the call centers is limited to 8 kHz, and the extracted features are often of low quality [28, 29, 30]. Affective expressions in other natural speech corpora have been induced by watching emotional videos [31, 32, 33], interacting with artificial listeners (humans or robots) [34, 35, 36, 37], participating in interviews with research teams [38, 39], or participating in either natural or artificial meetings [40].

Since the majority of emotional states are difficult to elicit in laboratory settings, the acquisition of spontaneous affective speech data is still an unresolved problem. While it is relatively easy to elicit joyful laughter by showing subjects clips from comedies, other states such as fear, anger, or love, are much more difficult to elicit [6].

Another major challenge in constructing emotional speech databases is the acquisition of ground truth. In many databases, it is even difficult for human subjects to label the emotion of the recorded utterances. Even with the databases using actors, human recognition accuracy was very low; in fact, it was only 67% for the Danish emotional database [41] with five emotional classes, and 65% for the Emotional Speech of Mandarin and Burmese Speaker database with six emotional classes [4, 42].

1.3 Baseline Features

Automatic speech recognition (ASR) has a longer history than affective computing, and its database and features are standardized; the Texas Instruments and Massachusetts Institute of Technology (TIMIT) speech and Wall Street Journal corpora are the most well-known, and the mel-frequency cepstral coefficients (MFCC's) and their derivatives are standardized optimal features. In the case of affect recognition, the optimal feature set has not yet been established [6]. In the past two decades, affective speech researchers have put

forth laborious effort to find relevant features. Although the optimal feature set has not yet been established, many researchers in emotion recognition employ the openSMILE toolkit as their primary feature extractor. The latest version of openSMILE extracts up to 6,373 acoustic features from speech signals. The openSMILE feature extractor provides energy, spectral, and voicing-related low-level descriptors, along with their statistical and regression measures [43, 44, 45].

In general, a larger number of features does not always result in better classification. It is important to reduce the dimensionality of the feature set, not only to speed up the classification process, but also to optimize classification performance. Feature projection algorithms are often employed for this reason.

Feature projection algorithms use statistical methods to reduce the dimension of the features by applying linear transformation. Two popular feature projection algorithms are principal components analysis (PCA) and linear discriminant analysis (LDA). The main difference between the two is that LDA finds the optimal transformation matrix that can be used to discriminate between different classes, whereas PCA finds the optimal orthogonal linear transformation matrix that preserves the subspace with the largest variance without paying any particular attention to the underlying class structure [46, 47]. In general, LDA outperforms PCA, since LDA deals directly with class discrimination [46].

1.4 Emotion Classifiers

For the last two decades, affective-computing researchers have focused their work on finding relevant features, and numerous studies are based on existing pattern recognition techniques such as support vector machines, Gaussian mixture models, hidden Markov models, and random forests. It was only recently that researchers started investigating novel classifiers to better model the unique nature of emotions. However, even these recent methods are not far from the fusion of existing algorithms. Some of the recently proposed methods with their unweighted accuracies (UWA) on emotion corpora are shown in Table 1. A brief

summary of each method is followed in the table.

Various studies have shown the benefits of different features and learning algorithms at the phonic or utterance level. The emotion is encoded at different levels of speech, and each timescale feature is complementary. Samuel Kim et al. extracted spectral and prosodic features at supra- and intra-frame-levels and employed a late fusion algorithm to produce the final classification decisions [48]. In their work, a Gaussian mixture model (GMM) was adopted to represent the distribution of the intra-frame features, and the k nearest neighborhood algorithm (k-NN) was chosen to model the supra-frame features. For the final decision, the weighted sum of likelihoods from the two classifiers was used. By tuning the weighting factor between the two classifiers, their proposed algorithm obtained 95% accuracy (unweighted) on the electromagnetic articulography (EMA) database. This database consists of 1,964 utterances collected from three speakers (one male and two females).

Table 1: Recently proposed emotion classifiers.

Authors	**Database**	**Classes**	**Classifier**	**UWA**
S. Kim et al. [48] in 2007	EMA Database [49]	neutral, angry	GMM + k-NN	95%
C. Lee et al. [50] in 2009	FAU-AEC Corpus [51]	neutral, positive, rest, emphatic, angry	hierarchical binary decision tree	42%
C. Wu et al. [52] in 2011	2 hour database	neutral, happy, sad, angry	SVM + GMM + MLP + MDT	80%
H. Meng et al. [15] in 2011	SEMAINE Corpus [37]	activation, expectation, dominance, valence	k-NN + HMM	53%
S. Ntalampiras et al. [53] in 2012	BERLIN Database [54]	neutral, happy, surprised, sad, angry, fearful	HMM + MLP	92%
A. Metallinou et al. [1] in 2012	IEMOCAP database [3]	valence (neg, neu, pos) and activation (low, med, high)	HMM + NN	(A)52% (V)65%

In the work by Chi-Chun Lee et al., they developed a hierarchical binary decision tree, where the top-level classification is performed on the easiest emotion recognition task [50].

The main idea of their work is to split the five-class problem into a set of two-class problems. Using either a general Bayesian logistic regression model or a support vector machine at each layer, their algorithm classifies two subsets of emotions. At the top layer, the algorithm classifies two subsets of emotions, {angry, emphatic} and {positive}, which they claim to be the easiest task. At the next layer, the algorithm distinguishes between {angry} and {emphatic}. They used the FAU-AEC database, which consists of children's emotional speech collected from fifty-one children interacting with an AIBO dog [55]. With the five emotional classes in the database, the unweighted recall was 42%. The algorithm improves the unweighted recall by 3% (absolute) compared with using a support vector machine.

Chung-Hsien Wu et al. proposed an emotion recognition approach based on multiple base-level classifiers using acoustic-prosodic features and semantic labels [52]. On the base level, Gaussian mixture models (GMM), support vector machines (SVM), and multi-layer perceptrons (MLP) are used for emotion classification, and a meta decision tree (MDT) is employed for the fusion of the three classifiers. The role of the MDT is to select the most promising classifier for acoustic-prosodic-based emotion recognition. The difference between an MDT and an ordinary decision tree is that the leaves in MDT specify which base-level classifier should be used instead of predicting the fusion probability of the emotional state directly. Moreover, they developed a semantic-based classifier, then integrated it with the MDT classifier for the final emotion classification. The authors used two Chinese dialogue corpora, which consist of four emotional classes with two hours of duration in total. Their algorithm achieved 81% average accuracy with the semantic labels, and 67% without the semantic labels.

Naturalistic emotional expressions change slowly as a person interacts with the environment. To model the temporal process of emotion, a fully connected HMM was employed by Tin Lay Nwe et al. [42]. Hongying Meng et al. expanded the idea, and proposed a multi-stage approach based on hidden Markov models. In the first stage, the method first predicts the four affective dimensions by the K-nearest neighbor algorithms. The method

then trains the HMMs based on the decision values from the first stage. Lastly, the method combines the multiple classifiers into another HMM in the third stage to boost the overall performance. The authors reported that the unweighted average recall was improved by 6% (absolute) compared to using a support vector machine.

Stavros Ntalampiras et al. [53] also employed HMMs to model the temporal behavior of emotions. After the distribution of the feature values with respect to each emotional category is approximated by HMMs, fusion is conducted using a multilayer perceptron algorithm. Using the BERLIN database, which consists of six emotional states, their algorithm has 92% accuracy.

Emotions are slowly varying states, so an angry utterance is more likely to be succeeded by one displaying anger rather than happiness. In the work by Angeliki Metallinou et al. [1], the authors proposed a neural network-based algorithm that takes into account an arbitrary amount of past and future audiovisual emotional expressions to recognize the current emotion of a speaker. At the lower level, HMMs are used to model each utterance as a sequence of audiovisual observations. At the higher level, neural networks are used to model an emotional conversation as a sequence of emotional utterances. Using the IEMOCAP database, which contains facial motion capture (MoCap) information as well as speech information, the authors performed 3-level classification on two affective dimensions, *activation* and *valence*. The unweighted accuracy on valence using audiovisual features was 65%, and it was 52% on activation.

1.5 Summary

The field of speech emotion recognition is emerging and its task is very challenging. One of the most difficult problems in the field is that emotion does not have a commonly agreed theoretical definition [4, 56]. According to Rafael Calvo [5], there are six different perspectives/theories to describe emotions: 1) emotions as expressions, 2) emotions as embodiments, 3) emotions in cognitive approaches, 4) emotions as a social construct, 5) emotions

in neuroscience, and 6) emotions as a psychological construct. Most of the current work in classification is limited to Darwin's theory on emotion. Many studies have focused on predicting emotional state based on people's vocal or facial expressions. Since the acquisition of spontaneous affective data in real-life settings is not an easy task, most emotional speech databases and their ground-truth labels are obtained in terms of "expressions."

Another important issue in affective computing is the labeling process of emotions. Different perspectives lead to different methods for labeling. For example, a person who is mourning the death of a spouse may express other transient emotions that will not last for more than a few minutes. In this case, researchers adopting Darwin's perspective on emotions will focus on the transient emotions that the person expresses in a short time, and researchers adopting cognitive approaches will argue that the true emotional state of the person should be analyzed after understanding his/her situation. No clear boundaries exist between these two temporal analyses with respect to long-term and transient emotion [4].

1.6 Organization

The objective of this research is to analyze affective expressions and to develop probabilistic models for multimodal affect recognition. This dissertation is centered on investigating relevant features using advanced signal processing techniques and designing a classifier fusion method for multimodal-multitemporal analyses. The classifier fusion algorithm is intended to integrate classifiers trained at various analysis lengths, as well as those obtained from other modalities, such as visual signals. Although this and similar research is limited by inconsistent subjective boundaries and differing theoretical definitions of emotion, a significant step toward successful emotion classification has been demonstrated; key to the progress has been via novel voice and visual features and a newly developed multimodal-multitemporal approach.

The dissertation is organized with three main topics. In Chapters 2 and 3, acoustic features for affect classification are discussed by presenting two novel feature sets extracted

from formants and multi-resolution spectral analysis. In Chapters 4 and 5, a novel classifier fusion method is introduced to model the multimodal-multitemporal nature of emotion. In Chapter 6, the techniques explored for emotion classification are used in practical dyadic plays between a child and an adult.

CHAPTER 2

FORMANT-BASED FEATURES FOR EMOTION CLASSIFICATION

As evidenced in the preceding chapter, many acoustic features have been used for emotion analysis. Among the most important of these are features that describe the pattern of resonances within the vocal tract, also known as *formants*. Vocal tract resonances, or formants, have received continued attention over past decades, and continue to play an important role in applications, such as automatic speaker identification [57], clinical depression diagnosis [58, 59], and children's speech therapy [60]. The shape of the vocal tract and its resonance descriptors are also important for emotion classification. Vivien Tartter showed that smiling raised the formant frequencies as well as the fundamental frequencies [61]. Robert Frick showed that threat was generally perceived due to a lowering of the formant frequencies caused by increases in the size of the vocal tract [62]. It has been also found that subjects under depression do not articulate voiced sounds with the same effort as in the neutral emotional state. The slackened, articulated speech from the subjects under depression has wider formant bandwidths than those in the neutral emotional state [63].

Although copious psychology and behavior literature has shown that articulatory characteristics in emotional speech are correlated with formants, their use in automatic emotion classification has been limited. Since formants highly depend on the target sounds, they have generally been analyzed across the same phonemes. If this constraint were to be relaxed, better emotion classification might be possible in utilizing formants. To this end, a novel method of representing formants is introduced in this chapter. Instead of using only the individual frequencies, amplitudes, and bandwidths of formants, their interrelations (e.g., differences and ratios) are also estimated. It is shown that their interrelations are less dependent on phonetic characteristic.

Most formant estimators are based on linear predictive coding, which often fails with

nasalized phonemes. Moreover, these estimators are vulnerable to environmental noise often rendering their estimates as unreliable features. A novel and more accurate formant estimation algorithm, whose features also produce better classification results, is introduced in this chapter.

2.1 Formant-based Feature Extraction using Linear Predictive Coding

The baseline feature set primarily consists of prosodic, spectral, and energy; their statistic, regression, and local minima/maxima related functionals produce the total number of the 1,941 features in the set [64]. As an extension of the spectral features in the baseline set, the formant-based features are extracted, and used in this study.

This section explores formant-based features extracted with linear predictive coding (LPC) analysis for emotion classification in the four affective dimensions, namely activation, expectation, dominance, and valence. The SEMAINE database [37] was used for word-level emotion classification. According to well-established psychological literature [65, 66, 64], activation is the individual's global feeling of dynamism or lethargy involving mental and physical activity. Expectation is an emotion involving pleasure and excitement in considering some expected or longed-for good event. The dominance dimension subsumes two related concepts, power and control. For instance, while both fear and anger are unpleasant emotions, anger is a dominant emotion, while fear is a submissive emotion. Valence is an individual's overall sense of weal or woe [64].

Many algorithms have been introduced to describe its resonances and cross-section areas during emotional speech production [63]. The formants, a representation of the vocal tract resonances, can be modeled using linear predictive coding. The first three formants were found using LPC, and the coefficients were used to extract features that describe their amplitudes, frequencies, and the bandwidths. To find the formant-based features, the speech signal was downsampled to 8kHz, and filtered with a pre-emphasis filter as in Eq.

(1). To boost the energy in the high frequency band where SNR is low, α was set to 0.97.

$$H(z) = 1 - \alpha z^{-1} \tag{1}$$

The filtered signal, with sampling frequency (fs), was divided into frames using 30 ms Hamming windows with 15 ms overlap. A 10-pole LPC analysis was performed on each frame, and the complex root pairs of the LPC polynomial were calculated. The angle ω_k and radius r_k of the pole correspond to the frequency and amplitude of the k^{th} formant. For $r \approx 1$, the radius of the pole is also related to the formant bandwidth BW_k as shown in (2) [67]. An example of the spectral envelope estimation of the vowel /o/ is depicted in Figure 1.

$$BW_k = g\left(\frac{1 - |r|}{\sqrt{|r|}}\right)\frac{fs}{2\pi} \qquad \text{,where } g = 2 \tag{2}$$

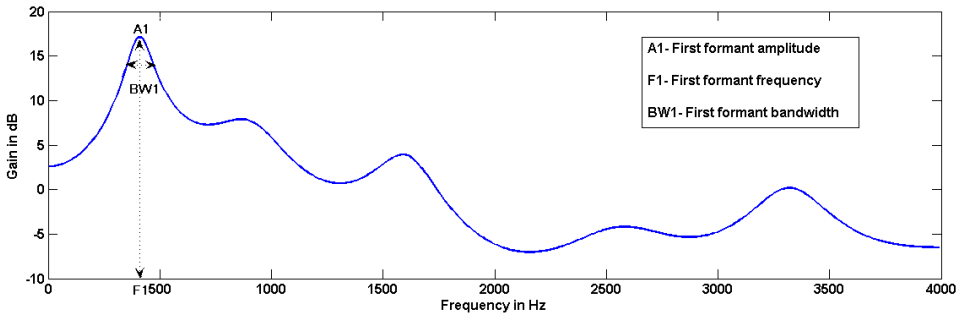


Figure 1: 10-pole LPC response of the vowel /o/.

After obtaining the first three formant frequencies, amplitudes, and bandwidths, their interrelations were described by the differences and the ratios shown in Table 2. The time series of these 18 features were then represented with 14 statistical measures and 4 regression measures as shown in Table 3. This produces 324 formant-related features.

2.1.1 Feature Selection

With the LPC-formant-based features, the total of 2,265 features was obtained. In general a larger number of the features does not always result in better classification. It is important to reduce the dimensionality of the feature set, not only to speed up the classification process,

Table 2: Formant-related low-level descriptors (LLDs).

Feature	Description
A1 ~ 3	First 3 formant amplitudes
F1 ~ 3	First 3 formant frequencies
BW1 ~ 3	First 3 formant bandwidths
A1 - A2	Difference between A1 and A2
A2 - A3	Difference between A2 and A3
A1 - A3	Difference between A1 and A3
F1 - F2	Difference between F1 and F2
F2 - F3	Difference between F2 and F3
F1 - F3	Difference between F1 and F3
BW1 / BW2	Ratio of BW1 to BW2
BW2 / BW3	Ratio of BW2 to BW3
BW1 / BW3	Ratio of BW1 to BW3

Table 3: List of statistical and regression measures for formant LLDs.

Type	Measure
Statistical measure	maximum, minimum, mean, standard deviation, kurtosis, skewness, flatness, 1^{st}, 2^{nd}, and 3^{rd} quartiles, interquartile range, 1^{st} and 99^{th} percentiles, and root mean square value
Regression measure	slope of linear regression, approximation error of linear regression, quadratic regression coefficient, and approximation error of quadratic regression

but also to optimize classification performance. For this purpose, three algorithms were compared. Two feature selection algorithms explored are based on the criteria of maximal relevance (MaxRel) and minimum-redundancy-maximal-relevancy (mRMR) [68]. The third algorithm was based on maximal average recall (MaxARC). The feature selections for all three algorithms are done in two stages. The first stage was to rank the features according to each algorithm's criterion, and the second stage was to wrap the optimal feature set using sequential forward selection.

2.1.1.1 Maximal Relevance

The maximal-relevance method computes the set of features S, consisting of m features. The features are ranked from 1 to m by computing the mutual information between the solitary feature x_i and class c. The resulting set of features S has the largest dependency on

the target class c as shown in Eq. (3).

$$\max D(S,c), \quad D = \frac{1}{|S|}\sum_{x_i \in S} I(x_i;c) \tag{3}$$

The mutual information between two random variables x_i and c is defined as,

$$I(x_i;c) = \sum\sum p(x_i,c)\log\frac{p(x_i,c)}{p(x_i)p(c)} \tag{4}$$

where $p(x_i,c)$ is the joint probability distribution function of x_i and c. The probability distribution functions of x_i and c are $p(x_i)$ and c, respectively.

2.1.1.2 Minimal-Redundancy-Maximal-Relevance

The minimal-redundancy-maximal-relevance method is a two-pronged approach to select the optimal set of features. The maximal-relevance in Eq. (3) is first computed and followed by the minimal-redundancy. The redundancy is computed by finding the mutual information between features as in Eq. (5).

$$\min \ R(S), \quad R = \frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i;x_j) \tag{5}$$

The mRMR criterion optimizes the difference between D and R, and the features are ranked by the values of ϕ.

$$\max \ \phi(D,R), \ \ \phi = D - R \tag{6}$$

It was shown that MaxRel and mRMR tend to perform better on the discrete features than the continuous ones [69], and it was recommended to discretize the features based on their mean values and standard deviations. Using two thresholds, mean ± one standard deviation, the continuous features were discretized into 3 states.

2.1.1.3 Maximum Average Recall

In the maximum average recall (MaxARC) method, the features are ranked according to the average recall rate as defined in Eq. (7).

$$\text{Average Recall Rate} = \frac{1}{2}\left(\frac{TP}{TP+FP} + \frac{TN}{TN+FN}\right) \tag{7}$$

In Eq. (7), *TP* stands for true positive, *FP* for false positive, *TN* for true negative and *FN* for false negative. Since the data set for this study was highly unbalanced, the average recall criterion was chosen as an alternative. The average recall of each individual feature was calculated over 10-fold cross-validation using a GMM classifier with 32 mixtures. Unlike the other two methods, this method is classifier dependent.

2.1.1.4 Sequential Forward Selection

Using only the training set of the corpus, the features were ranked from 1 to 1000 by the three algorithms described above. Since the m best features are not the best m features [68], it is necessary to find the optimal subset of the features. Given the subset of ranked 2165 features, a sequential forward selection algorithm was used to wrap the features by adding features one-by-one in rank order [70]. Starting with first ranked feature, the feature x_i was added to the subset and tested for the reduction in the error rate. The feature x_i that did not result in an improvement in the error rate was discarded, and the feature resulted in the error rate reduction was kept in the subset. The error rate was tested with a 10-fold cross-validation using the GMM classifier.

2.1.1.5 Evaluation

The three feature selection criteria were used for the four models developed for the affective dimensions as shown in Figure 2. It can be observed that the MaxARC criterion's result for activation is comparable to that of mRMR for most of the features with both of them performing better than MaxRel. MaxARC has the lowest error rate in expectation with more number of features. In dominance, there is a steep descent in the error rate in the first few iterations for MaxRel and mRMR. After the 16^{th} feature, MaxARC converges with mRMR while MaxRel reaches to the lowest error rate for dominance. In valence, MaxRel gives the lowest error rate. Thus, feature subsets selected by MaxRel give the lowest error rates for valence and dominance, and the selections of MaxARC give the lowest error rates for activation and expectation. There is no single winner for all four affective dimensions.

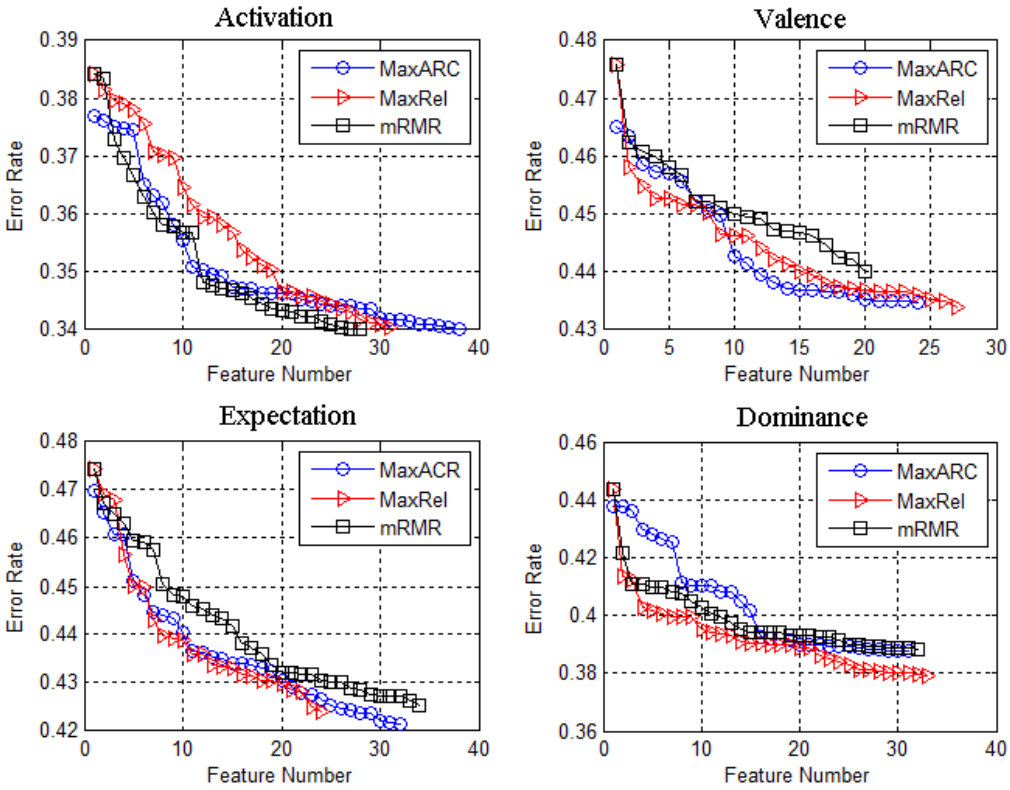


Figure 2: Error rate vs. number of features for MaxRel, mRMR, and MaxArc using a 10-fold cross-validation method on the training set.

2.2 Evaluation on Formant-based Features Extracted using LPC

The effect of formant-based features was investigated in two ways. First, the results from the feature selection criteria were used to study how a single formant feature improved the error rate when added in the subset. Second, 10-fold cross-validation was performed to compare before and after the formant-based features. Using the optimal feature subsets found in Section 2.1.1, the average recall as a function of feature number is plotted in Figure 3.

The highest ranked formant feature is "A3 50% quantile" for valence; it is ranked as the fourth most relevant feature. The formant feature that improves the average recall rate the most is "(A1-A3) 25% quantile"; it improves 0.2% recall rate. More information on the selected formant-based features in the subset are shown in Table 4.

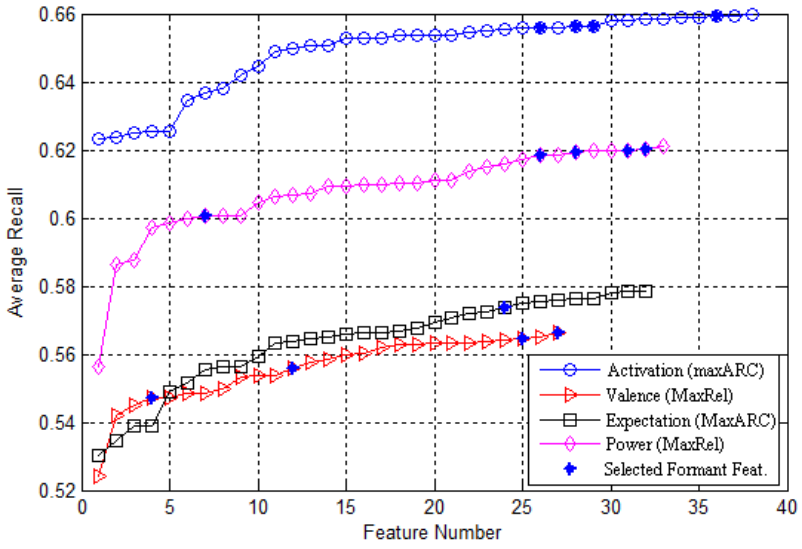


Figure 3: Average recall (unweighted accuracy) vs. number of features to indicate the formant-based features selected by the feature selection criteria in the training set using a 10-fold cross-validation method.

Table 4: Formant-based features selected using MaxARC, MaxRel and mRMR in the training set.

Activation MaxARC		Valence MaxRel		Expectation MaxARC		Dominance MaxRel	
Rank	Name	Rank	Name	Rank	Name	Rank	Name
26	F3 50% quantile	4	A3 50% quantile	24	F1 min	7	F3 50% quantile
28	BW2 1^{st} percentile	12	(A1-A3) 25% quantile			26	F1 flatness
29	BW2 min	25	A2 99^{th} percentile			28	(BW1/BW2) 25% quantile
36	(BW2/BW3) flatness	27	(BW1/BW3) min			31	BW2 25% quantile
						32	F1 linear reg. curve slope

For each affective dimension, classification was performed using the three feature selection criteria and a 10-fold cross-validation technique; this gives 30 classification trials on each affective dimension, and the average was taken as shown in Table 5.

Table 5: Average recall rate on the training set using the three selection criteria and a 10-fold cross-validation technique.

	Activation	Valence	Expectation	Dominance
Baseline	65.93	57.23	57.78	61.90
Formants	61.42	55.12	53.15	55.89
Baseline +Formants	66.14	57.42	57.82	62.25

2.2.1 Classification on the Development Set

The training set of the corpus was used to train the four affective dimensions with 32 Gaussian mixtures, and binary classification was done on the development set. Once again, the selected feature sets from Section 2.1.1 were used to test the development set. The results are shown in Table 6, and the best unweighted accuracy for each affective dimension is highlighted in bold. It is consistent with Figure 2 that the MaxARC feature set produced the best unweighted accuracy on the development set for activation, and mRMR feature set for dominance; however, the mRMR feature set produced the highest unweighted accuracy on the development set for valence and the MaxRel feature set for expectation.

Table 6: Classification results on the development set. ARC represents average recall rate while ACC stands for weighted accuracy. Base+Form is the combination of baseline and formant-based features.

	Activation		Valence		Expectation		Dominance	
	ARC	ACC	ARC	ACC	ARC	ACC	ARC	ACC
Base+Form MaxRel	64.40	64.35	51.31	54.31	**53.95**	60.67	**56.18**	63.96
Base+Form mRMR	63.42	63.65	**54.19**	53.36	53.08	59.04	53.97	61.48
Base+Form MaxARC	**65.62**	65.05	51.83	54.17	53.06	61.26	55.46	61.79

The possibility of including the formant-based features in the existing repertoire of baseline features for affective classification was explored. It was shown that 14 formant-based features were selected by the three feature selection algorithms; five formant frequency, three amplitude, and six bandwidth related features were selected. The inclusion

of formant-based features with the baseline features has shown an improvement in the average unweighted accuracy on the train set with a 10-fold cross-validation method compared to the average unweighted accuracy using only the baseline features for all the four affective dimensions.

2.3 Formant Frequency Tracking using Gaussian Mixtures

Most formant tracking algorithms are based on linear predictive coding (LPC). LPC-based algorithms produce a reasonable approximation during vowel-like sounds; however, LPC-based formant trackers often encounter problems with modeling speech elements, such as nasalized phonemes and give inconsistent results for bandwidth estimation [71]. The LPC analysis technique is based on the assumption that the speech production system is an all-pole filter, so in the case with the nasal and fricative sounds, the formant frequencies and bandwidths cannot be well estimated. Moreover, LPC-based techniques tend to underestimate the formant bandwidths of high-pitch voiced signals, because the harmonic spacing is too large to provide an adequate sampling of the spectral envelope [72].

In spite of their significance in classifying emotional speech, formants have not yet been widely adopted as acoustic features. To be useful acoustic features, John Holmes et al. suggest that formant frequencies be supplemented by general spectral shape information [73]. For this reason, a statistical method of estimating spectral parameters using a Gaussian mixture model (GMM) is investigated in this section.

The proposed formant tracking algorithm has its foundation in the work of Parham Zolfaghari and Tony Robinson in which Gaussian mixture distributions were fitted to discrete Fourier transform (DFT) magnitude spectra [74]. Zolfaghari and Robinson's work focused on finding maximum-likelihood (ML) estimates from a single DFT analysis frame. The proposed method extends this work by using the sequence of wideband spectra with MAP adaptation to refine the estimates. The hypothesis is that an accurate formant frequency estimation can be achieved by spectral modeling using a Gaussian mixture with

MAP-adaptation algorithm. By using MAP adaptation with a reasonable adaptation rate, the proposed method produces the smooth transitions from frame to frame. An efficient method for Gaussian parameter estimation is also done by utilizing DFT amplitudes.

2.3.1 Parameter Initialization by Expectation-Maximization

As discussed in [74], various problems are encountered when fitting Gaussian distributions directly to the DFT magnitude spectra. One of the major challenges is that the Gaussian fitting can pick a single harmonic and neglect adjacent harmonics. This particular problem occurs more often with female speakers, where fundamental frequencies are high and harmonics are widely spaced. As a possible solution, [74] suggested applying cepstral smoothing [75], which has the effect of removing the high-quefrency excitation components from the spectrum.

The proposed formant tracking algorithm first estimates Gaussian mixture parameters for a sequence of M wideband spectra, then uses a single spectrum in the middle of the sequence to re-estimate the parameters. To find the parameters in the sequence of spectra, the expectation-maximization (EM) algorithm is used. Maximum *a posteriori* (MAP) adaptation is then used to refine the estimates by adapting the parameters based on their neighbors. The EM algorithm applied to the sequence of spectra resolves the overfitting problem, and the MAP-adaptation algorithm produces accurate parameter estimates for the given frame with smooth transitions from frame to frame [76]. The process is depicted in Figure 4.

An expectation-maximization algorithm originally designed for finding maximum likelihood estimates of parameters in statistical models, therefore to estimate the GMM parameters from the spectrum, a probability density function must be formed from the spectrum [77, 74, 78]. To fit a set of Gaussians to a spectrum, Stuttle suggests to form a histogram from continuous bin probability functions [78]. The proposed approach is similar to the suggested method, but the spectrum amplitudes are directly used to obtain the estimates with less computational cost. The approach is described below.

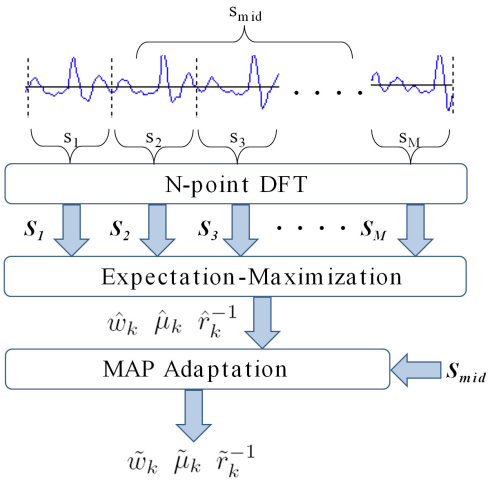


Figure 4: System overview of the proposed method with a sequence of M analysis frames.

A Gaussian mixture model is a weighted sum of K component Gaussian densities as follows:

$$g(x_t \mid \lambda) = \sum_{k=1}^{K} w_k \mathcal{N}(x_t \mid \mu_k, r_k), \tag{8}$$

where

$$\mathcal{N}(x_t \mid \mu_k, r_k) \propto |r_k|^{1/2} \exp\left[-\frac{1}{2}(x_t - \mu_k)' r_k (x_t - \mu_k)\right], \tag{9}$$

w_k is the weight of the k^{th} Gaussian density, μ_k is the mean, and r_k is the inverse of a D x D covariance matrix [77]. The goal is to estimate the Gaussian mixture parameters, λ in Eq. (10), such that the resulting curve best fits the aggregated DFT magnitude spectra.

$$\lambda = (w_1, ..., w_K, \mu_1, ..., \mu_K, r_1, ..., r_K). \tag{10}$$

The maximum-likelihood (ML) estimate of the parameters λ^*_{ML} in Eq. (11) can be obtained by the EM algorithm, where it finds the parameters iteratively, such that $g(\mathbf{x} \mid \lambda^{i+1}) > g(\mathbf{x} \mid \lambda^i)$ for each iteration i [77].

$$\lambda^*_{ML} = \underset{\lambda}{\operatorname{argmax}}\, g(\lambda \mid \mathbf{x}) = \underset{\lambda}{\operatorname{argmax}}\, g(\mathbf{x} \mid \lambda). \tag{11}$$

Since GMMs fit their curves to the distribution of the feature vector $\mathbf{x}$, the amplitudes of the N-point DFT spectra are discretized to form a histogram-like representation of their

corresponding frequencies as follows:

$$\mathbf{x} = \{f_1, ..., f_1, f_2, ..., f_2, ..., f_N, ..., f_N\}, \tag{12}$$

where the number of occurrences of f_n is its discretized amplitude a_n, and the length of $\mathbf{x}$ is $T = \sum_{n=1}^{N} a_n$. The EM algorithm first finds the *a posteriori* probability for the k^{th} Gaussian component with an initial λ as follows:

$$c_{kt} = \Pr(k \mid x_t, \lambda) = \frac{w_k \mathcal{N}(x_t|\mu_k, r_k)}{\sum_{l=1}^{K} w_l \mathcal{N}(x_t|\mu_l, r_l)}, \tag{13}$$

where it is evaluated over T values of x_t [78].

It is important to note that there are N unique values in $\mathbf{x}$; therefore, the *a posteriori* probability can be simplified as follows:

$$c_{kn} = \frac{w_k a_n \mathcal{N}(f_n|\mu_k, r_k)}{\sum_{l=1}^{K} w_l a_n \mathcal{N}(f_n|\mu_l, r_l)}. \tag{14}$$

The probabilistic count for x_t belonging to the Gaussian component k is defined as follows:

$$c_k = \sum_{t=1}^{T} c_{kt} = \sum_{n=1}^{N} c_{kn}. \tag{15}$$

On each EM iteration, the likelihood value increases monotonically when the parameters are re-estimated as follows [79]:

$$\hat{w}_k = \frac{c_k}{T} = \frac{c_k}{\sum_{n=1}^{N} a_n}. \tag{16}$$

$$\hat{\mu}_k = \frac{\sum_{t=1}^{T} c_{kt} x_t}{c_k} = \frac{\sum_{n=1}^{N} c_{kn} f_n}{c_k}. \tag{17}$$

$$\hat{r}_k^{-1} = \frac{\sum_{t=1}^{T} c_{kt}(x_t - \hat{\mu}_k)(x_t - \hat{\mu}_k)^{'}}{c_k} = \frac{\sum_{n=1}^{N} c_{kn}(f_n - \hat{\mu}_k)(f_n - \hat{\mu}_k)^{'}}{c_k}. \tag{18}$$

Eq. (11) is equivalent to finding the ML estimate of the Gaussian mixture parameters for N frequencies with their corresponding DFT magnitudes as weights, shown in Eq. (19).

$$\lambda_{ML}^{*} = \underset{\lambda}{\operatorname{argmax}} \prod_{n=1}^{N} \sum_{k=1}^{K} w_k a_n \mathcal{N}(f_n \mid \mu_k, r_k). \tag{19}$$

By utilizing c_{kn} and the amplitude a_n, the computation is more efficient when compared to the direct EM algorithm, where the Gaussian mixture curve is fitted to the distribution of $\mathbf{x}$.

2.3.2 Parameter Estimation by MAP Adaptation

The main difference between MAP and ML estimation is that in the former, the parameters, λ, are treated as random variables [80] so that the distribution function, $h(\lambda)$, is no longer assumed to be constant. MAP estimation is defined as

$$\lambda^*_{MAP} = \underset{\lambda}{\operatorname{argmax}}\, g(\lambda \mid \mathbf{x}) = \underset{\lambda}{\operatorname{argmax}}\, g(\mathbf{x} \mid \lambda)h(\lambda). \tag{20}$$

Similar to the ML-EM algorithm, the MAP estimation first finds the expectations of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}^2$ with the newly acquired data $\tilde{\mathbf{x}}$ as follows:

$$E_k(\tilde{\mathbf{x}}) = \frac{\sum_{n=1}^{N} \tilde{c}_{kn}\tilde{f}_n}{\tilde{c}_k}, \tag{21}$$

$$E_k(\tilde{\mathbf{x}}^2) = \frac{\sum_{n=1}^{N} \tilde{c}_{kn}\tilde{f}_n^2}{\tilde{c}_k}. \tag{22}$$

In the present case, $\tilde{\mathbf{x}}$ is the vector of the frequencies of the middle spectrum, S_{mid}, and their corresponding amplitudes are used when calculating $\tilde{c}_{kn}$, as in Eq. (14).

The ML-EM parameters are then adapted to the DFT magnitude spectrum of the middle frame with an adaptation rate ρ, as follows:

$$\tilde{w}_k = \rho\frac{\tilde{c}_k}{\sum_{n=1}^{N} \tilde{a}_n} + (1-\rho)\hat{w}_k, \tag{23}$$

$$\tilde{\mu}_k = \rho E_k(\tilde{\mathbf{x}}) + (1-\rho)\hat{\mu}_k, \tag{24}$$

$$\tilde{r}_k^{-1} = \rho E_k(\tilde{\mathbf{x}}^2) + (1-\rho)(\hat{r}_k^{-1} + \hat{\mu}_k^2) - \tilde{\mu}_k^2, \tag{25}$$

where $\hat{w}_k$, $\hat{\mu}_k$, and $\hat{r}_k^{-1}$ are the old estimates obtained in the previous EM steps. The new estimates are combined with the old parameters obtained from the previous EM algorithm with the adaptation rate ρ chosen between 0 and 1. When ρ is close to 0, the new parameter estimates are adapted to the new data $\tilde{\mathbf{x}}$ at a fast rate; when ρ is close to 1, the parameter adaptation is slower.

In typical speaker adaptation models used in speech recognition, ρ is defined as $\frac{\tilde{c}_k}{\tilde{c}_k+\gamma}$ where γ is determined empirically. In the proposed method, if S_{mid} is similar to its neighbors, slow adaptation is preferred for a smooth transition between frames. If S_{mid} is not

similar to its neighbors, the method emphasizes new parameters and de-emphasizes the old ones. Therefore, a similarity measure is used as the adaptation rate, which is measured using cross-correlation coefficients as follows:

$$\rho = \frac{\sum_{m=1}^{M} \mathrm{corr}(S_{mid}, S_m)}{M}, \tag{26}$$

where M is the number of neighboring frames employed.

The EM algorithm is iterative and well-known for training Gaussian mixtures. The novelty in the current analysis relates to the inclusion of adjacent frames of spectra in the iteration, giving better results. The formant frequencies are obtained from the means of the Gaussian mixtures, $\tilde{\mu}_k$, and the formant amplitudes are their weights, $\tilde{w}_k$. The formant bandwidth estimates are proportional to the standard deviation, $\sqrt{\tilde{r}_k^{-1}}$.

2.3.3 Formant Frequency Tracking Results

Testing was performed using a vocal tract resonance (VTR) database which is a representative subset of the TIMIT speech corpus. The VTR database consists of 516 utterances with three manually labeled formants (F1, F2, and F3) for each 10 ms [81]. For the proposed method, "six" adjacent 5 ms-analysis windows were used during the EM stage, and one 20 ms-analysis window for MAP adaptation. The original speech signals were down-sampled from 8,000 Hz to 4,000 Hz, and four Gaussian mixtures were fitted to DFT magnitudes from 0 Hz to 4,000 Hz. Two examples of the estimated formant tracks along with the hand-labeled ground truth are depicted in Figure 5.

The resulting formant trajectories are smooth and well-behaved over voiced frames, but the trajectories tend to have higher estimated frequencies than the ground truth over unvoiced frames, especially for affricates. This tendency is shown at 1.5 and 1.9 seconds of Figure 5(b). Since the affricates have high energy at high frequency, this tendency is expected.

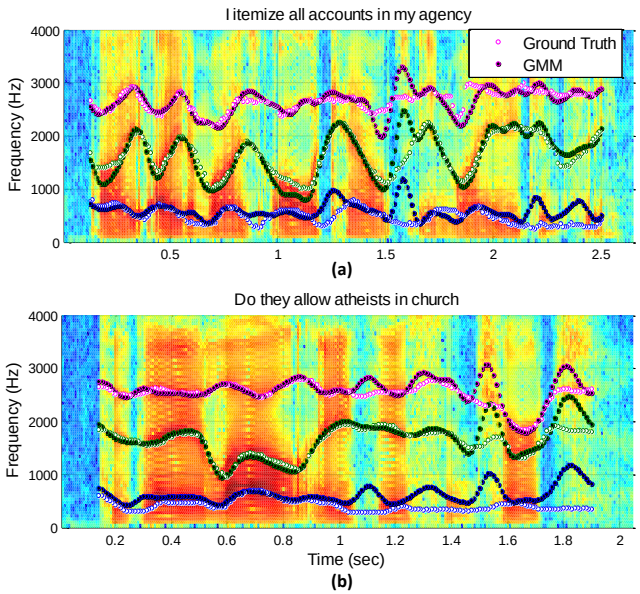


Figure 5: Spectrograms of two utterances with formants estimated by the proposed method (filled circles) and the hand-labeled ground truth (empty circles).

2.3.3.1 Experiment I: Comparison with Other Methods

For comparison, the first four formants were extracted using three other methods. The first method is a formant tracker proposed by Zolfaghari, in which Gaussian mixture distributions are fitted to magnitude spectra with cepstral smoothing (GMM+cep in Table 7). The second one is a widely used and highly developed speech-processing tool, PRAAT, whose algorithm is based on linear predictive analysis. The third one is a formant tracker based on a time-varying adaptive filter bank described by Mustafa and Bruce [82]. For the formant tracker proposed by Zolfaghari (GMM+cep), 30 ms-analysis windows with a 10 ms-frame interval were used. For PRAAT, 30 ms-analysis windows with a 5 ms-frame interval were used. A 9th-order lowpass Butterworth filter with a cutoff frequency of 10 Hz was then applied for smoothing the extracted tracks of PRAAT. The lowpass filter reduced the root mean square error (RMSE) of PRAAT by 4%. For the formant tracker proposed by Mustafa and Bruce (MB), 30 ms-analysis windows were used. Using hand-labeled formant tracks of the VTR database as the ground truth, the RMSE was calculated for the first three formants for every 10 ms. The average errors over all frames for six phonetic classes were

calculated and are shown in Table 7.

Table 7: RMSE of the proposed method (GMM+MAP), GMM+cep, PRAAT, and MB for the first three formants.

Phonetic	GMM+MAP			GMM+cep			PRAAT			MB		
class	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3
Vowel	75	118	128	96	125	151	82	127	146	101	158	208
Semi-vowel/glide	88	138	177	118	166	201	105	147	181	130	193	303
Nasal	95	234	193	112	213	195	132	258	207	183	230	206
Fricative	238	282	298	335	327	348	405	298	311	201	211	233
Affricate	409	347	303	607	366	353	610	347	308	200	346	191
Stop	164	211	236	254	253	291	314	246	265	168	219	247
Overall per formant	127	178	190	178	198	223	196	194	208	141	190	229

In all cases, the proposed method outperforms PRAAT. This is especially true for first formant frequency estimation. When compared to GMM+cep, it is shown that the inclusion of MAP adaptation has significantly improved the formant tracking results. The proposed method outperforms GMM+cep in all cases, except for the second formant frequency estimation of *nasals*.

Mustafa and Bruce's method performs the best for the fricative and affricate-phonetic classes when it is compared to the other three methods; however, their method does not perform as well as the others for the sonorant-phonetic classes (vowels, semivowels, and nasals). Since vocal tract resonances are fairly well defined for voiced signals, the average error of the sonorant-phonetic classes are relatively smaller than that of the occlusive-phonetic classes (fricatives, affricates, and stops).

2.3.3.2 Experiment II: Evaluation in Noisy Environments

The proposed method is also evaluated at different SNRs, and compared with the results from PRAAT for the sonorant-phonetic classes. The original clean signals were degraded by white and babble noise from the NOISEX-92 database at various SNRs. The RMSE of the first three formants using the two methods at different SNRs is depicted in Figure 6. When the signals were degraded by white noise, the proposed method outperforms PRAAT at all tested SNRs as shown in Figure 6 (a).

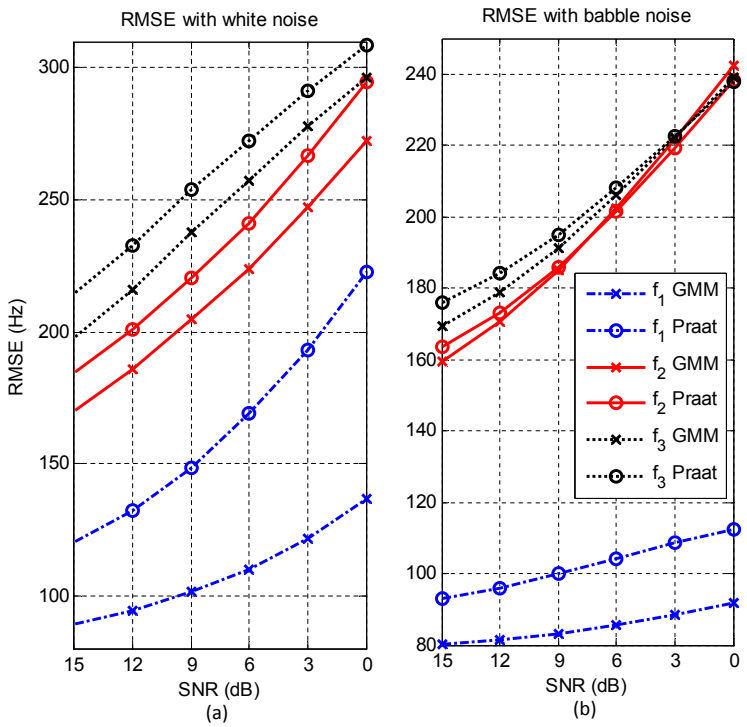


Figure 6: RMSE at various SNRs with (a) white noise and (b) babble for the sonorant-phonetic classes.

The RMSE of the first formant significantly improved with the proposed method. With babble noise, the proposed method still outperforms PRAAT at the SNRs higher than 3dB. With SNRs lower than 3dB, errors of the proposed method for the second formant are slightly higher than those of PRAAT, although, the proposed method still outperforms PRAAT for estimating the first formant at all tested SNRs as shown in Figure 6 (b).

To summarize the results, the proposed method significantly reduces the errors when it is compared to PRAAT's LPC-based algorithm. With clean signals, the proposed method improves the relative errors for all the phonetic classes by 35%, 8%, and 9% for the first three formants, respectively. The proposed method was also evaluated with white noise and babble at various SNRs, and it was shown that the proposed method is less vulnerable to noisy environments when compared to PRAAT.

Although Gaussian mixtures do not conform to classical speech production models, they are able to fit speech spectral data robustly. Further, they offer a natural framework

for adaptation and smoothing that can be used advantageously in tracking time-varying formants.

The inclusion of MAP adaptation shows significant improvement over a similar GMM-based formant tracking framework.When compared to an LPC-based method and one based on speech-motivated filter banks, certain patterns and trends are observed. The proposed method is superior to both competing systems for voiced segments, where formant tracks are continuous and smooth from frame to frame. The implementation of MAP adaptation successfully models this behavior in the sonorant-phonetic classes.

For unvoiced segments, an advantage is still maintained over the LPC-based tracker, and is competitive with the filter-bank-based system. Of particular note is the relatively poor performance over the relatively infrequent affricate class, where formants themselves are of questionable use and interpretation (similar for plosives). However, the aggregate accuracy over all classes indicates a strong preference for the proposed method.

For noisy data, it is not surprising that the new approach, which does not rely on a parametric speech model, such as LPC, is a more accurate method. This is particularly so with white noise. When the LPC order remains fixed, and a noise signal of sufficient amplitude is present, an LPC-based method completely misses formants in order to model noise peaks [83].

2.4 Evaluation of Formant-Based Features Extracted Using GMM

2.4.1 Geneva Multimodal Emotion Portrayals Corpus

For formant-based feature evaluation, the Geneva multimodal emotion portrayals (GEMEP) corpus was used. The GEMEP corpus was collected from 10 subjects (5 males and 5 females) by University of Geneva and Austrian Research Institute [27]. The ten subjects are professional, French-speaking actors. They portrayed 18 emotional expressions under the direction of a professional stage director. The authors of the corpus address the pros and cons of acted emotion expressions. One of the main shortcomings is that acted

portrayals are often exaggerated and not sufficiently spontaneous. Despite the lack of spontaneity, acted portrayals offer numerous advantages. Emotions underlying acted portrayals can be defined precisely [26], and they offer the possibility of recording variable expressions by the same individuals. With "natural" expressions, only a few emotional reactions can be recorded for a given individual. To collect more spontaneous data, the authors employed the Stanislavski or Method acting technique. Hence, expressions emerge more spontaneously from the emotional state the actors have tried to produce, and "believable" emotions can be drawn [26, 27, 3].

The actors were requested to improvise interactions with the director; the director's goal was to encourage the actors to steer "genuine" emotional expressions without predefined visual/acoustic realization of expressions (e.g., no instruction to smile, to frown, or to shout). More than 7000 such emotion portrayals representing 18 emotions were collected. Three trained psychology students, who served as research assistants, assessed the technical quality of the recordings and the aptitude of the actors to convey the intended emotional impression. Based on their assessments, 126 portrayals per actor were selected to provide 1260 emotional utterances of 18 categorical emotions in 2 affective dimensions: *activation* and *valence* [27]. Since six of the 18 emotional categories are sparse, they were excluded from evaluation. The remaining 12 classes include *amusement, anxiety, cold anger, despair, elation, hot anger, interest, panic fear, pleasure, pride, relief*, and *sadness*.

In Section 2.2, the inclusion of formant-based features for word-level emotion classification showed significant improvement for the SEMAINE database [37]. This previous work had used formant-based features extracted by an LPC-based algorithm. The goal of this section is twofold: 1) to compare the features from two formant tracking algorithms: GMM+MAP and PRAAT, and 2) to evaluate the inclusion of formant-based features for emotion classification in the GEMEP database.

2.4.2 Classification Results

Since there were no phonetic or word timestamps in the GEMEP database, a 400 ms analysis-window length was chosen to parse the 1,260 utterances in the database. The resulting number of segments is 12,512. After obtaining the 324 formant-based features using GMM+MAP and PRAAT, as described in Section 2.1, a 10-fold cross-validation technique was performed on two affective dimensions and 12 categories of emotion, with results shown in Table 8. For classifying 12 categories of emotion with SVMs, the "one-against-one" approach was used [84, 85]. If k is the number of classes, then $k(k-1)/2$ binary classifiers are constructed, and each classifier trains data from two classes. In the classification stage, the approach employs a voting strategy, where the final decision is made by choosing the class label with the most votes from individual SVMs.

Table 8: Unweighted average accuracy over 10 folds using the formant-based features analyzed with 400-ms windows using the proposed method (GMM+MAP) and PRAAT in the GEMEP database.

Classifier	Activation		Valence		12 Categories	
	GMM+MAP	PRAAT	GMM+MAP	PRAAT	GMM+MAP	PRAAT
GMM (M=4)	77.8	65.8	64.2	62.7	24.6	20.3
GMM (M=8)	77.5	62.7	63.2	63.0	26.4	21.7
SVM (C=0.1)	79.4	78.3	63.1	62.0	32.2	29.5
SVM (C=1)	80.4	78.9	64.8	63.7	36.9	32.5

Gaussian mixture models and support vector machines were used with different configurations. For GMMs, 4 and 8 mixtures were used, and 324 features were projected using linear discriminant analysis (LDA) before the training. For SVM training, the complexity parameter, C, was set either to 0.1 or 1. As shown in Table 8, the formant-based features extracted by the GMM+MAP method resulted in higher accuracy than those extracted using PRAAT.

For all cases, the SVM with C=1, produced the highest unweighted accuracy. For significance testing, the p-value between GMM+MAP and PRAAT was calculated for a

paired t-test. For *activation*, the UWA was improved by 1.5 percentage points, with a p-value less than 0.025. For *valence*, the UWA was improved by 1.1 percentage points, with a p-value less than 0.05. Finally, for classification into 12 categories, UWA was improved by 4.4 percentage points with a p-value less than 0.01. For all cases, the GMM+MAP method improved the UWA significantly. Also note that the improved formant-tracking algorithm (GMM+MAP) produced improved detection and classification accuracies.

The inclusion of formant-based features in the baseline feature set was also investigated. The baseline features were calculated using the popular feature extractor, openSMILE, which produced 6,373 acoustic features from speech signals [43]. The openSMILE feature extractor provides energy, spectral, and voicing-related low-level descriptors, along with their statistical and regression measures [43, 44, 45]. In spite of their significance in classifying emotional speech, formants are not included in the openSMILE feature set. 10-fold cross-validation was performed to study the effect of the formant-based features, with results shown in Table 9.

Table 9: Unweighted average accuracy over 10 folds using SVMs (C=1) with the combination of formant-based and baseline features analyzed with 400-ms windows.

	Activation	Valence	12 Categories
Baseline	81.3	68.8	39.3
Baseline+Formants (PRAAT)	80.0	70.4	40.7
Baseline+Formants (GMM+MAP)	83.1	72.2	42.3

As can be seen in Table 9, the inclusion of formant-based features using the proposed method (GMM+MAP) along with the baseline feature set statistically significantly improved the UWA. The results suggest that the formant-based features contain additional information on affect not revealed by the baseline features. For activation, GMM+MAP improved the UWA by 1.8 percentage points with a p-value less than 0.05. The amount of improvement was highest in the valence dimension, where it was 3.4 percentage points, with a p-value less than 0.0025. For classifying 12 categories of emotion, the inclusion of formant features using GMM+MAP improved the UWA by 3.0 percentage points with

a p-value less than 0.01. When the two formant feature extractors were compared, the combination of the baseline and GMM+MAP features produced 3.1, 1.6, and 1.6 percentage points higher UWA for activation, valence, and 12 categories, respectively. When a paired t-test was performed for the two formant feature extractors, the improvement made by GMM+MAP was statistically significant, with a p-value less than 0.05 for all cases. The confusion matrices using the combined features are shown in Tables 10 and 11 for the two emotional dimensions and 12 categories, respectively. Each row represents the instances in an actual class normalized by the total number of instances, and each column represents the normalized instances in a predicted class.

For activation and valence, the confusion matrices indicate that the true positive rates and true negative rates are well-balanced, and the classifiers do not favor a single class. In the case of the 12-way classifier, the diagonal terms (correctly classified) are almost always highest except for the emotional category, "panic fear," where it was classified as "hot anger" more often than its own class. The lexical definitions of these two classes are clearly different, but they both belong to a "high activation and negative valence" state. This pair of emotional categories is a good example that shows the world of emotion is not simply two-dimensional [66]. Although, the two categories are very similar in the activation and valence dimensions, their differences can be accentuated when a third emotional dimension, "dominance," is introduced. Anger is a dominant emotion, whereas fear is a submissive emotion. It is known that classifying the dominant states is relatively difficult when only the speech modality is used. The current GEMEP corpus does not include the labels for dominance, and further investigation on this dimension will be carried out using another corpus called IEMOCAP in Chapter 5.

Table 10: Confusion matrices of classifying two levels of activation and valence, using the combination of baseline and formant-based features analyzed with 400-ms windows. The classification was done at the utterance level using SVMs (C=1).

Activation			Valence		
	high'	low'		pos'	neg'
high	81.8%	18.2%	pos	73.2%	26.8%
low	15.6%	84.4%	neg	28.8%	71.2%

rows: ground truth; columns: hypothesis

Table 11: Confusion matrix of classifying 12 categories of emotion, using the combination of baseline and formant-based features analyzed with 400-ms windows. The classification was done at the utterance level using SVMs (C=1).

	amusement	anxiety	cold anger	despair	elation	hot anger	interest	panic fear	pleasure	pride	relief	sadness
amusement	**69**%	3%	2%	14%	5%	3%	0%	2%	1%	2%	0%	1%
anxiety	8%	**29**%	13%	9%	2%	3%	6%	0%	11%	3%	5%	11%
cold anger	3%	11%	**38**%	6%	2%	1%	10%	0%	10%	5%	5%	9%
despair	20%	6%	3%	**39**%	5%	7%	2%	2%	5%	2%	2%	7%
elation	22%	3%	4%	11%	**43**%	5%	1%	2%	1%	6%	0%	1%
hot anger	7%	3%	3%	13%	7%	**55**%	0%	1%	0%	9%	0%	0%
interest	0%	4%	8%	2%	1%	1%	**36**%	0%	22%	2%	8%	15%
panic fear	12%	5%	3%	18%	11%	26%	1%	**18**%	0%	1%	2%	3%
pleasure	1%	5%	6%	3%	0%	1%	12%	0%	**57**%	1%	6%	8%
pride	8%	5%	14%	6%	6%	12%	4%	0%	7%	**32**%	2%	2%
relief	4%	8%	11%	4%	0%	1%	9%	0%	17%	2%	**38**%	6%
sadness	2%	7%	5%	3%	0%	0%	10%	0%	18%	0%	3%	**52**%

rows: ground truth; columns: hypothesis

2.5 Conclusion

In this chapter, the use of formant-based features for emotion classification was investigated and shown to be valuable. The first tests used a LPC-based formant tracking algorithm. After estimating the first three formant frequencies, amplitudes, and bandwidths, their interrelations were described by differences and ratios. In total, 324 formant-related features were extracted from 18 interrelations of formants, whose time-series representations were described by 14 statistical measures and 4 regression measures. Using the SEMAINE database, binary classification was performed at the word level over four affective dimensions. The results clearly show that the formants contain information on affect.

Since LPC-based formant estimators often encounter problems with modeling speech elements such as, nasalized phonemes and give inconsistent results for bandwidth estimation, a novel formant tracker was also introduced to better model the formants and spectral properties. The novel tracker estimates spectral parameters using Gaussian mixtures and a MAP adaptation algorithm to refine the estimates. Although Gaussian mixtures do not conform to classical speech production models, they are able to fit speech spectral data robustly. Further, they offer a natural framework for adaptation and smoothing, which can be used advantageously in tracking time-varying formants. The inclusion of MAP adaptation shows significant improvement over a similar GMM-based formant-tracking framework. When compared to an LPC-based method and one based on speech-motivated filter banks, certain patterns and trends are observed. The proposed method is superior to both competing systems for voiced segments, where formant tracks are continuous and smooth from frame to frame. The implementation of MAP adaptation successfully models this behavior in the sonorant-phonetic classes.

To evaluate the GMM+MAP method as feature extractor, formant-based features were extracted using the GEMEP corpus. For comparison, the features were also extracted using a LPC-based algorithm, PRAAT. When the formant-based features were evaluated for classifying two levels of activation and valence, GMM+MAP produced 1.5 and 1.1 percentage

points, respectively, higher UWAs than those of PRAAT. For classifying 12 categories of emotion, the UWA was improved by 4.4 percentage points when the formant-based features were extracted using GMM+MAP. In all cases, the classification results were improved statistically significantly, and it is plausible to conclude that formant-features extracted by a better formant estimator produce better classification results.

The inclusion of formant-based features in the baseline feature set was also explored. For activation, 1.8 percentage points of improvement in UWA was observed. For valence, the UWA was improved by 3.4 percentage points. For 12-way classification, the inclusion of formant-based features extracted using GMM+MAP improved the UWA by 3.0 percentage points. In all cases, the improvements made by the inclusion of formant-based features were statistically significant. The results strongly suggest that formant-based features contain additional information on emotion that the baseline features do not include.

CHAPTER 3

SPECTRAL FEATURE EXTRACTION USING MULTI-RESOLUTION SINUSOIDAL TRANSFORM CODING

Recent studies have shown that spectral features contain rich information about expressivity and emotion [86]. However, most of the work does not go beyond analyzing the mel-frequency cepstral coefficients (MFCC's) and their derivatives. Multi-resolution sinusoidal transform coding (MRSTC) will be explored in this chapter to analyze additional spectral properties. Because of MRSTC's high precision in representing spectral features, including preservation of high frequency content not present in the MFCC's, additional resolving power might be present. As an example, MRSTC can reveal whisper voices, which consist of less conspicuous harmonics than normal voices, and are good indicators of *grief* [8].

Typically, the harmonic-to-inharmonic ratio is calculated in a discrete Fourier transform (DFT) domain. However, speech signals are quasi-harmonic, and the harmonics in high-frequency bands may not be well estimated. A great strength of MRSTC is that it represents important auditory characteristics by performing several DFTs of differing lengths on the same data. The rationale behind MRSTC is that low-frequency signals are well defined in long analysis windows, whereas high-frequency signals require short analysis windows.

Moreover, the peak locations in MRSTC can be used to find the degree of pitch perturbation. Such pitch perturbation, or jitter, is defined as the period-to-period variability of the pitch period. The presence of pitch jitter can be observed in a low-frequency band with wider-peak bandwidths, and the ratio between harmonic peak amplitudes and their neighboring inharmonic peak amplitudes is noticeably lower in comparison with normally voiced speech signals. In this chapter, qualitative and quantitative analyses on the use of spectral features are investigated, with applications to affect classification.

3.1 Application of Spectral Features to Affect Classification

3.1.1 Feature Extraction using MRSTC

Prior to feature extraction, all speech signals were filtered with a high frequency pre-emphasis filter as in Eq. (1). As was done for the formant-based features, α was set to 0.97 to boost the energy in the high frequency band, where SNR is often low. After pre-emphasis, the signals were divided into frames using 32-ms Hamming windows with 16 ms overlap, and the root-mean-square (RMS) value was calculated for each frame. Imposing a RMS threshold resulted in the silent frames before and after each utterance being discarded prior to feature extraction.

Classical sinusoidal transform coding (STC) has been successfully used in many speech and audio applications [87, 88]. In continuous time, STC models the input signal as sum of AM-FM sinusoids:

$$s(t) = \sum_{k=1}^{K} A_k(t) \cos(\theta_k(t)), \tag{27}$$

where

$$\theta_k(t) = \Omega_k(t) + \psi_k(t). \tag{28}$$

The excitation phase $\Omega_k(t)$ is $(t - t_o)2\pi f_k$, where f_k is the k^{th} harmonic frequency, and t_o is the onset time. $\psi(t)$ is the system phase offset. In discrete short-time, the sinusoidal signals at m^{th} analysis frame are represented as follows:

$$s(n; m) = \sum_{k=1}^{K} A_k(m) \cos[(n - n_o(m))2\pi f_k(m) + \psi(m)], \tag{29}$$

where $A_k(m)$ and $f_k(m)$ at m^{th} frame are estimated by the spectral envelope estimation vocoder (SEEVOC) peak-picking routine operating in the frequency domain [89]. SEEVOC searches sequentially for the largest spectral amplitudes and depends on the average pitch. It first searches for the largest peak, A_1 at f_1 in the interval $[\frac{f_o}{2}, \frac{3f_o}{2}]$, then searches for the largest peak in the next interval $[f_1 + \frac{f_o}{2}, f_1 + \frac{3f_o}{2}]$. The process is continued until the edge of the speech bandwidth is reached [89, 90]. An example of peak selection is shown in Figure 7.

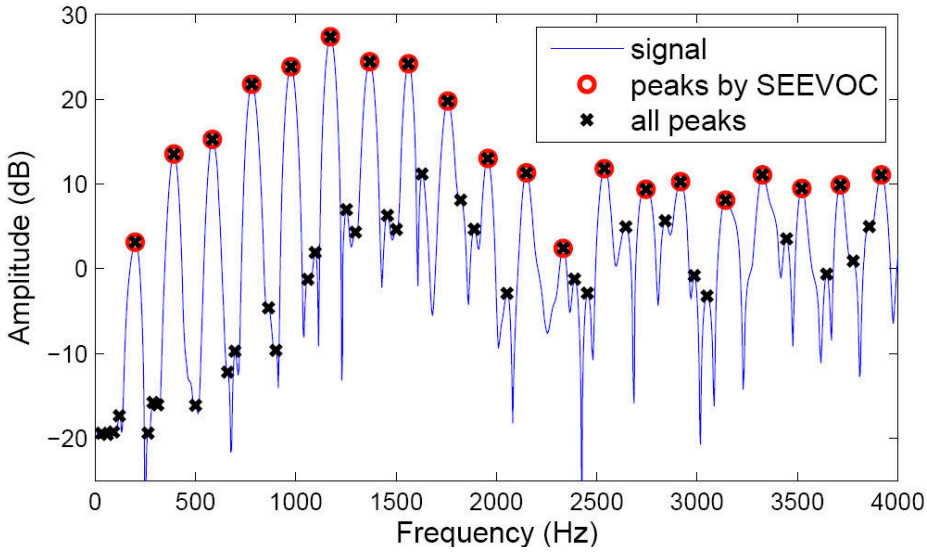


Figure 7: Harmonic peaks selected by SEEVOC peak-picking routine.

The speech signal is quasi-harmonic and non-stationary; its harmonic peaks are not perfectly equally-spaced by the fundamental frequency, and frequency content changes with time. Hence, a short-time Fourier transform (STFT) is usually used to analyze speech. In general, a 20-ms to 40-ms long analysis window is applied to speech signals to calculate the STFT. However, with a fixed window length, the subtle variations in pitch and harmonics, especially in high frequency bands, can be lost.

To capture these subtle changes in pitch and magnitude, a multi-resolution sinusoidal transform coding (MRSTC) method is employed. MRSTC uses wavelet-like analysis, where lower frequency components are calculated over a greater analysis window length, and higher frequency components are estimated with a shorter window length [2] as shown in Fig 8. A discrete wavelet tiling for a four-band MRSTC is shown in the figure, where the lowest frequency band is analyzed with a 32-ms window, and the highest frequency band is analyzed with an 8-ms window.

For scaling in the multi-resolution scheme, Daubechies' 10 wavelet (db10) was used to design the low-pass filters. The highpass filters were designed with quadrature mirror filters [91, 92]. After filtering, the signal was downsampled by a factor of 2 as shown in Figure 9. The sampling frequency of the corpus is 16 kHz, and the four sub-bands used in this work

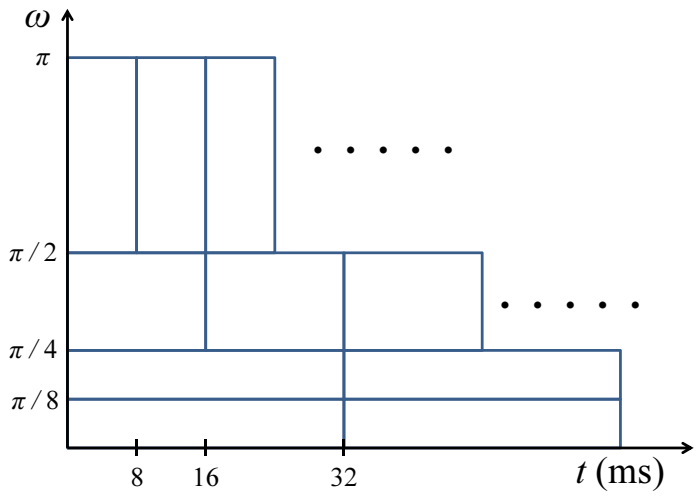


Figure 8: Discrete wavelet tiling for four-band MRSTC [2].

have the frequency ranges: [0 1000), [1000 2000), [2000 4000), and [4000 8000). The two lower sub-bands were analyzed with a 32-ms window while the third and fourth sub-bands were analyzed with a 16-ms and 8-ms window, respectively. The sinusoidal components, A_k and f_k, were found by using the SEEVOC peak-picking routine in each sub-band.

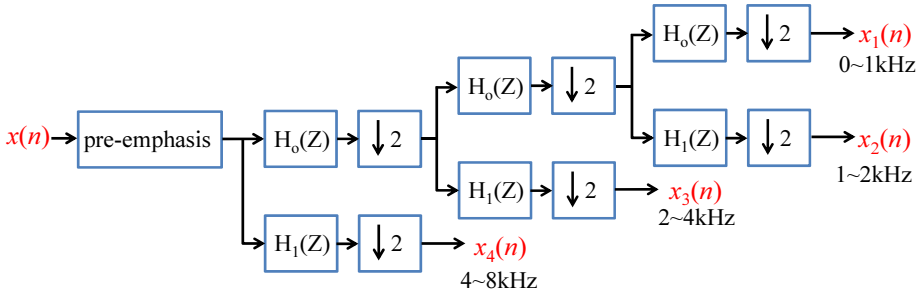


Figure 9: A quadrature mirror filter analysis bank arranged for four-band MRSTC [2].

Pitch perturbation, jitter, is defined as the period-to-period variability of the pitch period. The presence of pitch jitter can be observed in a low frequency band with wide peak bandwidths, and the ratio between harmonic peak amplitudes and their neighboring inharmonic peak amplitudes is noticeably lower when compared to normally voiced speech signals. Similar phenomena can be also observed when noise is added to the signal as shown in Figure 10. For demonstration purposes, a normally voiced signal received artificial jitter of 3% of the pitch, and white Gaussian noise (WGN) was added to this signal.

After noise addition, the SNR for the signal was roughly 7dB.

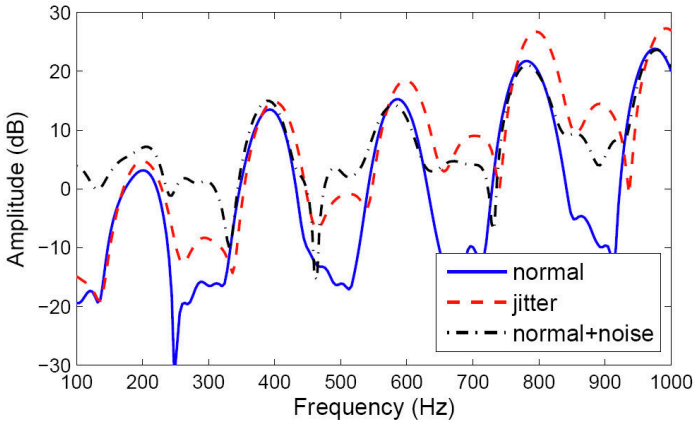


Figure 10: STFT of normal, normal+jitter, and normal+noise signals in a low frequency band.

Regardless of the source of perturbations, pitch jitter or additive noise, the harmonic-to-inharmonic peak amplitude ratio (HIR) is a useful measure for intelligibility, and is calculated as follows:

$$\text{HIR} = \frac{L}{K}\frac{\sum_{k=1}^{K} A_k}{\sum_{l=1}^{L} P_l}, \tag{30}$$

where A_k is the k^{th} harmonic amplitude in Eq. (29), and K is the total number of harmonics in a specified bandwidth. These amplitudes were found by the SEEVOC peak-picking routine while L inharmonic amplitudes, P_l's, were found by a typical peak-picking method. A decrease in the HIR is a good indication of degradation in voice quality due to a decrease of harmonic structure or an increase of additive noise in the source signal.

The autocorrelation of the short-time Fourier transform, $R_{FF}(f_{lag})$, was also calculated. In this autocorrelation domain, the signal with jitter is more distinguishable from the normal and normal+noise signals. As shown in Figure 11, the signal with jitter has more peaks than the other two signals. Because of the flat power spectral density, WGN does not affect the shape of the autocorrelation function, $R_{FF}(f_{lag})$ very much.

The average peak-to-peak distances in $R_{FF}(f_{lag})$ can serve as pitch estimates, but for signals with jitter, those peak distances are often shortened due to their wide, but distinct, inharmonic peaks in the STFT domain. This phenomenon motivates measuring the average

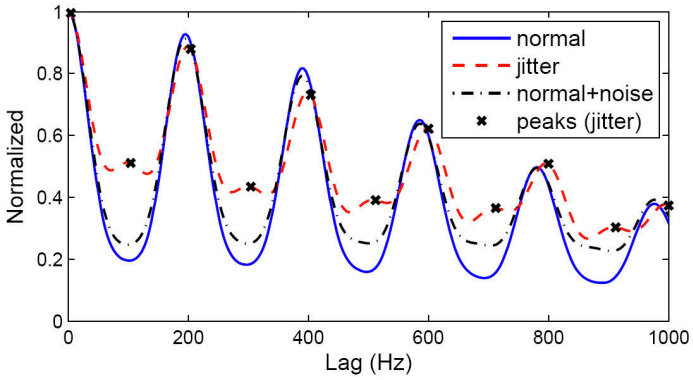


Figure 11: $R_{FF}(f_{lag})$ of normal, normal+jitter, and normal+noise signals.

peak-to-peak distance over the pitch in the STFT and $R_{FF}(f_{lag})$ domains as shown in Eq. (31). Along with the mean value, the standard deviation of $\Delta\frac{\mathbf{f}}{f_o}$ is also measured.

$$\text{Mean}(\Delta\frac{\mathbf{f}}{f_o}) = \frac{1}{K-1}\sum_{k=1}^{K-1}\frac{f_{k+1}-f_k}{f_o}. \tag{31}$$

At each analysis frame, the 9 statistical measures, marked with * in Table 12, were calculated for both the STFT and $R_{FF}(f_{lag})$ signals as low-level descriptors (LLDs). Moreover, all the measurements were analyzed in the four sub-bands to produce 92 LLDs per frame (23 LLDs $\times$ 4 bands). For utterance-level classification, the time series of all the LLDs were represented with the 14 statistical measures shown in Table 12. In total, 1,288 (23 LLDs $\times$ 4 bands $\times$ 14 measures) features were extracted in the MRSTC scheme, as shown in Figure 12.

Table 12: List of statistical measures for MRSTC feature extraction.

num.	description	num.	description
1	maximum	7	flatness
2	minimum	8~10	1^{st}, 2^{nd}, & 3^{rd} quartiles*
3	mean*	11	interquartile range*
4	standard deviation*	12~13	1^{st} & 99^{th} percentiles*
5	kurtosis*	14	RMS value
6	skewness		

*: measures applied to log (FFT) and the autocorrelation of STFT, $R_{FF}(f_{lag})$.

The baseline features were extracted using Technical University of Munich's (TUM) open-source feature extractor, also known as openSMILE [43]. The tool extracts statistical

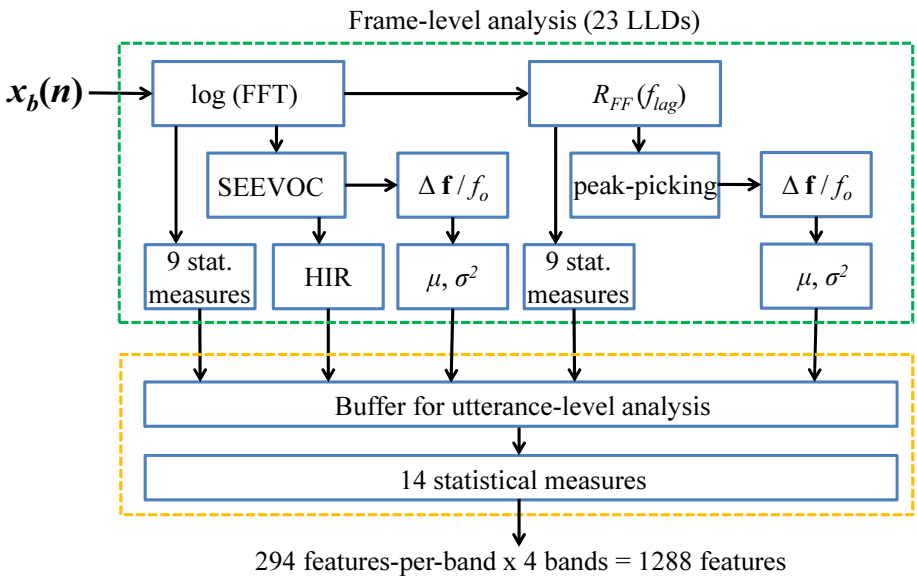


Figure 12: The proposed feature extraction method overview.

measures, regression measures, and linear prediction (LP) coefficients applied to energy, spectral, and voicing-related low-level descriptors (LLDs).

3.1.2 Classification Results

The GEMEP corpus was again used to evaluate the new spectral features extracted using the MRSTC scheme. 1,288 spectral features, 324 formant-based features, and 6,373 baseline features were extracted at both the 400-ms and utterance levels. In Section 2.4, the formant-based features, when evaluated only at the 400-ms level, showed that the features extracted using a better formant tracking algorithm (GMM+MAP) were much more beneficial in classifying two levels of activation and valence as well as the 12 categories of emotion than those using a traditional formant tracking algorithm (LPC).

In this section, the combinations of the three feature sets are evaluated at both the 400-ms and utterance level. The utterance-level decisions of the 400-ms level SVMs were made by using a majority vote method as in Section 2.4, and the utterance-level classifiers were again trained and tested with SVMs with a linear kernel and a cost parameter of 1. The individual feature sets and all their possible combinations were evaluated as shown in Table 13.

Table 13: Unweighted average accuracy over 10 folds using SVMs (C=1) with the combinations of baseline, formant, and spectral feature sets extracted at the 400-ms and utterance levels.

Feature sets	Activation		Valence	
	400 ms	Utt	400 ms	Utt
B	81.3%	82.2%	68.8%	76.9%
F	80.4%	76.5%	64.8%	63.4%
S	80.7%	81.4%	67.2%	69.5%
B+F	83.1%	83.0%	72.2%	77.6%
B+S	83.7%	85.1%	74.8%	79.7%
F+S	81.8%	79.2%	68.8%	71.1%
B+F+S	83.9%	85.3%	76.5%	80.2%

B: baseline features.
F: formant-based features extracted using GMM+MAP.
S: spectral features extracted using MRSTC.

When the individual feature sets were compared to one another, no significant differences were observed between the spectral and baseline feature sets in the activation dimension at either level. One sign of high activation is a tendency to raise the vocal intensity which can be characterized by spectral energy. Since the new spectral feature set includes the logarithmic amplitudes in four multi-resolution subbands, its performance in the activation dimension is as effective as the baseline feature set, where spectral energy is measured using 26 RASTA filters [93]. The formant-based feature set is comparably as good as the other two when analyzed at 400 ms, but its performance significantly degraded when analyzed at the utterance level. In the case of the formant-based features, the first three formant amplitudes are good indicators for measuring energy in time-varying resonant filters for a short period of time such as a phoneme or word-level. However, for a long duration analysis, the formants can appear at any frequency location, and their measurements (amplitudes, frequencies, and bandwidths) may lack coherence. For this reason, the formant-based features generally produce better classification results when analyzed with a short window length.

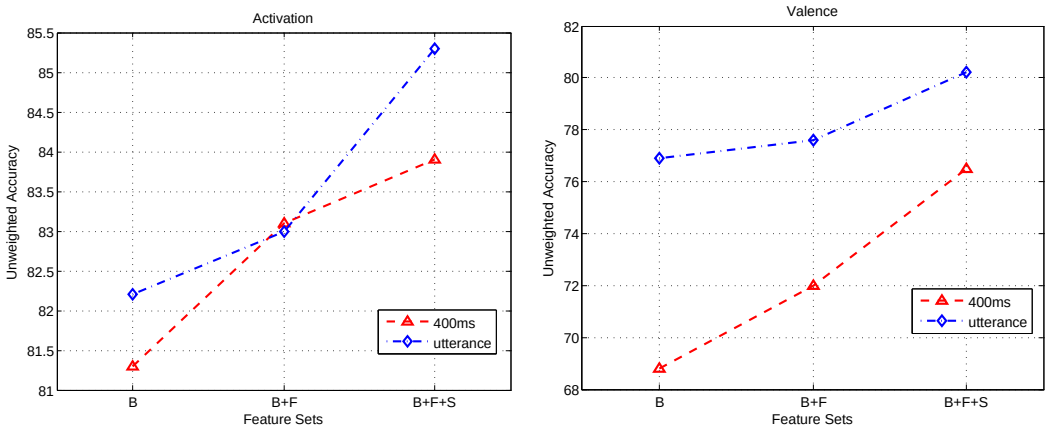
In the valence dimension, the baseline feature set is superior to other two feature sets. It is generally accepted that MFCC-related features, whose primary goal is to represent the

envelope of the spectrum, play the most important role in valence classification [94]. The new spectral feature set mainly describes the spectral peak-to-peak distances rather than shape of the spectrum. However, its classification accuracy is only 1.6 percentage points below the baseline feature set.

The goal of investigating new features is not to replace the existing baseline features, but to combine and to explore whether the new features contain any new relevant information on data. All possible combinations of the three feature sets were tested. As shown in Table 13, the highest classification accuracies are obtained when all the feature sets are combined for use in all cases. As discussed in Section 2.4, the combination of the baseline and formant feature sets (B+F), statistically significantly improved the classification accuracies in both the activation and valence dimensions when analyzed at the 400-ms level. Unfortunately, the effect of formant-based features is relatively low at utterance-level analysis. For activation, the unweighted accuracy was improved by 0.8 percentage points with p-value of only 0.059, and was improved by 0.7 percentage points with p-value of only 0.065. As discussed earlier, the statistical and regression measures of formants for a longer time period (utterance level), may lack coherence and lose emotional information.

The combination of the baseline and spectral features (B+S) improved the accuracy by 3.5 percentage points on average with a p-value less than 0.001. The highest improvement made was 6 percentage points in the valence dimension analyzed at the 400-ms level. The baseline feature set includes both spectral energy features and voice quality features (shimmer and jitter). The results indicate that the new spectral features extracted using a multi-resolution approach do contain important relevant information on emotional speech that the baseline features do to explain.

When the formant and spectral features were combined (F+S) with no baseline features, the classification results were as good as those using the baseline feature set for the 400-ms analysis. The number of combined features (F+S) is much smaller (1,288+324) than that of baseline features (6,373). This combination even outperformed the baseline



(a) Activation (b) Valence

Figure 13: Unweighted accuracy using SVMs (C=1) with baseline, formant, and spectral feature sets extracted at the 400-ms and utterance levels for (a) activation and (b) valence dimensions.

features (by a small amount) for the activation dimension analyzed at the 400-ms level. In contrast, the baseline feature set was much superior for the utterance-level analysis. The results indicate that these features are in a complementary relation, and they can benefit one another. As expected, the highest accuracies were obtained when all the three feature sets were combined in all cases. To better show the impact of the formant and spectral features, classification accuracies (UWA) are plotted in Figure 13.

As discussed earlier, the impact of the formant-based features (B to B+F) is relatively higher at the 400-ms analysis than the utterance-level analysis. The spectral feature set continues to improve the accuracies after the combination of the baseline and formant feature sets. The normalized confusion matrices for the final feature set (B+F+S) are shown in Table 14. The confusion matrices show that the true positive rates and the true negative rates are well-balanced and do not favor a single class.

3.2 Conclusion

Speech signals are quasi-harmonic, and the harmonics in high-frequency bands may not be well estimated when analyzed with a traditional discrete Fourier transform with a fixed

Table 14: Confusion matrices of classifying two levels of activation and valence, using the combination of baseline, formant-based, and MRSTC features analyzed with 400-ms and utterance levels. The classification was done at the utterance level using SVMs (C=1).

Activation						Valence					
400 ms			utterance			400 ms			utterance		
	high'	low'		high'	low'		pos'	neg'		pos'	neg'
high	81.5%	18.5%	high	81.3%	18.7%	pos	77.9%	22.1%	pos	80.4%	19.6%
low	13.7%	86.3%	low	10.5%	89.5%	neg	24.9%	75.1%	neg	20.0%	80%

rows: ground truth; columns: hypothesis

window length. A great strength of MRSTC is that it represents important auditory characteristics by performing several DFTs of differing lengths on the same data. In this chapter, a novel method for extracting spectral features in the MRSTC domain was introduced. To evaluate the effects of the spectral features extracted using the MRSTC scheme, the combinations of the three feature sets (formant, spectral, and baseline) were trained and tested using SVMs. All the features were extracted at both the 400 ms and utterance levels for evaluation.

First, each feature set was tested separately. The classification results indicate that the spectral feature set alone is as effective as the baseline feature set for activation classification. When the spectral and baseline feature sets were combined, the classification accuracies for valence and activation were statistically significantly improved by 3.5 percentage points on average. The results indicate that the new spectral features extracted using the multi-resolution approach do contain important relevant information on emotional speech that the baseline features do not explain.

In Chapter 2, formant-based features were extracted at the word level using the SEMAINE database, and extracted with 400-ms analysis windows using the GEMEP database. In this chapter, formant-based features were also extracted from utterance-level analysis and evaluated. The results show that the formant-based features are more effective when extracted with short analysis windows (400 ms) than when analyzed at the utterance level. However, the highest classification accuracy was obtained when all the three feature

sets were combined regardless of the length of analysis. The results indicate that these feature sets are in a complementary relation, and they benefit from one another in emotion classification.

The typical STFT-based algorithms, which use a fixed analysis window, may fail to characterize harmonic characteristics in a broad frequency band. The multi-resolution approach enables better representation of spectral and harmonic characteristics by using longer windows in lower-frequency bands, and shorter windows in higher-frequency bands. The experimental results clearly show that the features derived from the multi-resolution spectral domains are useful for emotion classification. By combining the new features with the baseline features, where no multi-resolution approach was considered, significant improvements in unweighted accuracy were observed.

CHAPTER 4

EMOTIONAL SPEECH ANALYSIS AT VARIOUS TEMPORAL LENGTHS

Different emotional characteristics can be observed at different timescales regardless of the modality of data [95]. At the phrase level, it has been shown that, in general, average pitch and intensity are higher with the emotional state of *hot anger* than with other states [96]. At the phonemic level, spectral tilt and formant frequency amplitudes are significantly different when the same phoneme is analyzed for different states [97, 7]. Similarly, jitter and shimmer measurements for pitch periods over 30-ms analysis windows are useful for detection of activation [98]. The work in this chapter focuses on analyzing acoustic-prosodic features at various temporal lengths.

Since emotional characteristics cannot all be modeled at fixed analysis frame sizes [95], it is reasonable to analyze speech with multiple analysis window lengths. However, combining multi-timescale features into a single representation is a challenging problem, and it becomes even more challenging with multimodal data. One of the issues with using multiple analysis window lengths is the asynchrony of the feature representations. A straightforward method to resolve the problem is to implement a late fusion algorithm, which combines the scores of multiple classifiers and trains on them to yield a final classification decision. Boosting is one of the most common late-fusion algorithms. However, most boosting-based algorithms require the classifiers to have the same decision level for fusion. Another simple approach is a linear weighted fusion (LWF) which combines classifiers by using a weighted sum of confidence measures at the decision level [10]. A serious limitation of LWF is that it often fails to model complex heterogeneous data.

A novel fusion algorithm, whose inputs are represented as multi-dimensional binary sequences resulting from individual classifiers, is introduced in this section. By binarizing the outputs of classifiers, synchronization of classifiers becomes a straightforward process,

and by employing a spectral clustering algorithm, the fusion method becomes capable of modeling heterogeneous data. The hypothesis is that an emotion classifier or detector can improve its performance when speech is analyzed at different timescales with fusion before a final classification decision.

4.1 Classifier Fusion with Binary Matrices

The proposed fusion method uses multiple classifiers trained at P temporal lengths, where each classifier is trained separately to classify M classes. The method first finds J clusters at each temporal length (p) using a Gaussian mixture model defined as follows:

$$g(x_t \mid \lambda) = \sum_{j=1}^{J} w_j \mathcal{N}(x_t \mid \mu_j, \sigma_j), \tag{32}$$

w_j is the weight of the j^{th} Gaussian mixture, μ_j is its mean, and σ_j is its D x D covariance matrix [77].

The maximum-likelihood (ML) estimate of the parameters is obtained by the EM algorithm, where it optimizes the parameters iteratively [77]. The goal is to find the cluster members that belong to each of the J Gaussian mixtures. Each data point of class m becomes a member of the j^{th} cluster when its likelihood is greater than a threshold τ_j, and its membership is represented by a binary number as follows:

$$\mathbf{A}_{m,p}(j,n) = \begin{cases} 1 & \text{if } \mathcal{N}(\mathbf{x}_n \mid \mu_j, \sigma_j) \geq \tau_j, \\ 0 & \text{otherwise.} \end{cases} \tag{33}$$

After obtaining all the binary matrices for all P analysis lengths, they are merged into one matrix either by up-sampling or down-sampling the individual matrices so that they all have the equal size. For class m, the resulting binary matrix, $\mathbf{A}_m$, is defined as follows:

$$\mathbf{A}_m(k,n) = [\mathbf{a}_m^1, \mathbf{a}_m^2, ..., \mathbf{a}_m^{N_m}], \tag{34}$$

where N_m is the number of instances in class m. The matrix $\mathbf{A}_m$ consists of N_m column vectors. Each column vector $\mathbf{a}_m^n$ has K elements which corresponds to the number of Gaussian

clusters (J) multiplies by the number of analysis lengths (P) as shown in the example in Figure 14 which shows an example of the $\mathbf{A}_m$ matrix with 2 analysis lengths and 4 Gaussian clusters after merging.

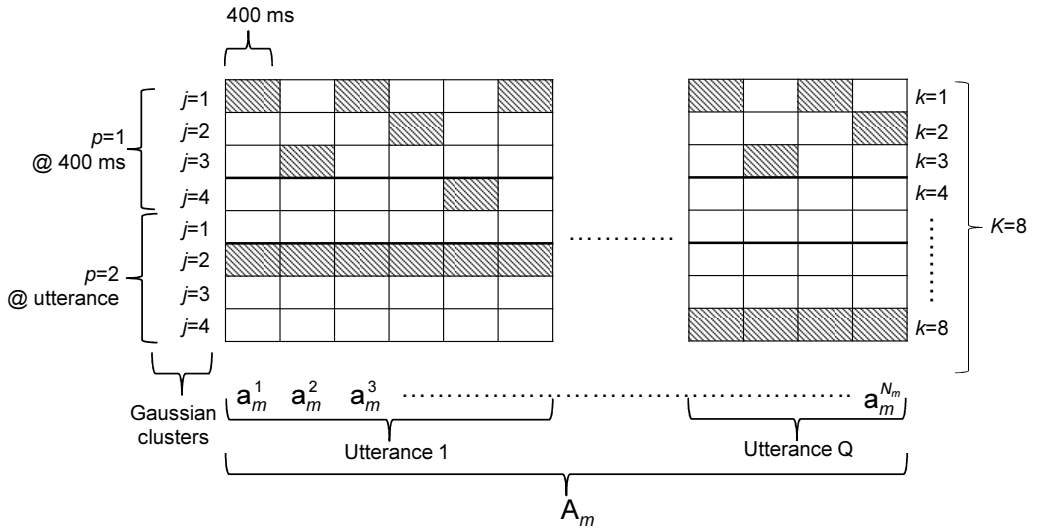


Figure 14: A graphical example of $\mathbf{A}_m$ matrix with 2 analysis lengths and 4 Gaussian clusters.

Because of the heterogeneity of the data, using one probabilistic model to fit the data is probably not adequate in many cases. A mixture model or clustering method is preferred when working with real-world data. One of the most common methods to train a binary matrix is a multi-variate Bernoulli mixture model. Since the binary matrix $\mathbf{A}_m$ is constructed with outputs from classifiers, using the weights based on the accuracy of the classifiers is useful for modeling or clustering data for training and testing.

When each row of the $\mathbf{A}_m$ matrix is viewed as the output sequence of the k-th detector, the posterior probabilities of the detector denoted as $\mathbf{w}_m^+(k)$ and $\mathbf{w}_m^-(k)$ can be calculated using the binary matrix $\mathbf{A}_m$ as follows:

$$\begin{aligned}\mathbf{w}_m^+(k) = \Pr(m|\mathbf{a}(k)=1) &= \frac{\Pr(\mathbf{a}(k)=1|m)\Pr(m)}{\sum_{i=1}^{M}\Pr(\mathbf{a}(k)=1|m_i)} \\ &= \frac{\sum_{n=1}^{N_m}\mathbf{A}_m(k,n)/N_m}{\sum_{i=1}^{M}\sum_{l=1}^{N_i}\mathbf{A}_i(k,l)/N_i},\end{aligned} \tag{35}$$

and with the negation of $\mathbf{A}$, the negative weight vector (the posterior probability of class m

given $\mathbf{a}(k) = 0$) is

$$\begin{aligned} \mathbf{w}_m^-(k) = \Pr(m|\mathbf{a}(k) = 0) &= \frac{\Pr(\mathbf{a}(k) = 0|m)\Pr(m)}{\sum_{i=0}^{M} \Pr(\mathbf{a}(k) = 0|m_i)} \\ &= \frac{\sum_{n=1}^{N_m} \bar{\mathbf{A}}_m(k,n)/N_m}{\sum_{i=1}^{M}\sum_{l=1}^{N_i} \bar{\mathbf{A}}_i(k,l)/N_i}. \end{aligned} \tag{36}$$

Note that $\Pr(\mathbf{a}(k)|m)$ is the sample mean of the data in class m which is also the maximum likelihood estimation for the Bernoulli distribution of each class m. Also note the equal priors are applied for all the classes, and the summation of $\mathbf{w}_m^+(k)$ and the summation of $\mathbf{w}_m^-(k)$ across M classes are equal to 1.

The goal is to cluster the binary data $\mathbf{a}_m$, whose dimension is $K \times 1$, into C clusters for training. To cluster the data, a similarity or distance matrix is first calculated. Using the binary column vectors, $\mathbf{a}_m^n$, the similarity between two instances can be measured by many different criteria. Since some of the k-dimensions are more relevant to the class m than the others, a dimension-weighted similarity matrix is constructed as follows:

$$\begin{aligned} \mathbf{S}_m(n,l) = &\left(\mathbf{a}_m^n \odot \mathbf{a}_m^l\right)^T \left|\mathbf{w}_m^+ - \frac{1}{M}\right| \\ &+ \left(\bar{\mathbf{a}}_m^n \odot \bar{\mathbf{a}}_m^l\right)^T \left|\mathbf{w}_m^- - \frac{1}{M}\right|, \end{aligned} \tag{37}$$

where $\odot$ denotes the Hadamard product. The similarity matrix is based on the inner-product similarity measure described in [99], and it is extended by imposing weights on both the positive and negative matches accordingly. The level of chance in classifying M classes, $1/M$, is subtracted from $\mathbf{w}_m^+$ and $\mathbf{w}_m^-$ in each dimension. The weights are chosen such that the similarity measure gets higher values when two instances match in higher discriminating dimensions than lower ones.

For example, consider a two-dimensional problem with two classes, where $\mathbf{w}_1^+$ is $[0.99\ \ 0.51]^T$, and $\mathbf{w}_1^-$ is $[0.2\ \ 0.48]^T$. For the second class, $\mathbf{w}_2^+$ and $\mathbf{w}_2^-$ are $(\mathbf{1} - \mathbf{w}_1^+)$ and $(\mathbf{1} - \mathbf{w}_1^-)$, respectively. Clearly, the first dimension (detector) is more discriminable than is the second one. The second dimension is almost meaningless and barely at the level of chance. When clustering the data into two clusters, grouping [0 0] with [0 1] and grouping

[1 0] with [1 1] are more beneficial than grouping [0 0] with [1 0] and grouping [0 1] with [1 1], because the first dimension is preserved. A spectral clustering algorithm is employed to this end.

4.1.1 Spectral Clustering using a Similarity Matrix

Many methods exist for clustering data from a similarity matrix. One well-known method is spectral clustering [100, 101, 102]. Its advantages include the fact that it does not make strong assumptions on the statistics of the clusters so it is very flexible with fitting data, and it is also simple to implement by using standard linear algebra methods [103].

For each class m, the spectral clustering method uses the similarity matrix, $\mathbf{S_m}$, to cluster N_m data points $\mathbf{a}_m^1$,..., $\mathbf{a}_m^{N_m}$ into C clusters. It first computes the Laplacian matrix, $\mathbf{L_m}$, as follows:

$$\mathbf{L}_m = \mathbf{I} - \mathbf{D}_m^{1/2}\mathbf{S}_m\mathbf{D}_m^{-1/2}, \tag{38}$$

where $\mathbf{D}_m$ is a diagonal matrix as shown in Eq. (39).

$$\mathbf{D}_m(i,i) = \sum_{j=1}^{N_m} \mathbf{S}_m(i,j). \tag{39}$$

With $\mathbf{L_m}$, the algorithm finds its first C eigenvectors to construct a matrix $\mathbf{V_m} \in \mathbf{R}^{N_m \times C}$, whose columns are the C eigenvectors [102]. After normalizing $\mathbf{V_m}$ such that each row has a unit length as in Eq. (40), it uses a k-means algorithm to cluster N_m rows into C groups [102].

$$\mathbf{U}_m(i,j) = \frac{\mathbf{V}(i,j)}{\sqrt{\sum_{r=1}^{C} \mathbf{V}^2(i,r)}}, i = 1,...,N_m, j = 1,...,C. \tag{40}$$

Consider the same example in Section 4.1, where $\mathbf{w}_1^+$ is $[0.99 \;\; 0.51]^T$, and $\mathbf{w}_1^-$ is $[0.2 \;\; 0.48]^T$ with two detectors and two classes. The resulting $\mathbf{U}$ matrix of the example is plotted in Figure 15.

As explained in Section 4.1, the binary vectors are clustered in a manner to preserve the output from the first detector that is more discriminant than the second one.

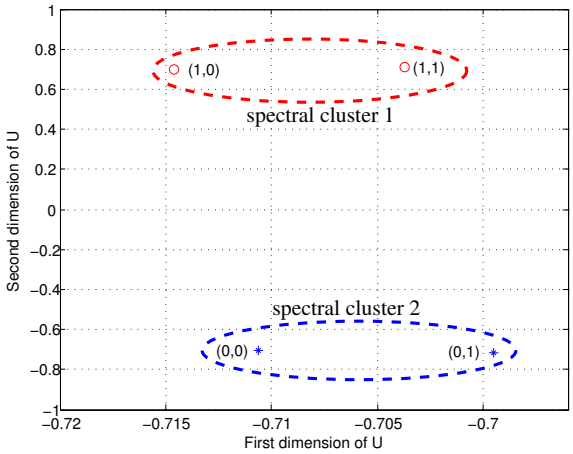


Figure 15: An example of the spectral clustering with four dimension-weighted binary vectors.

4.1.2 Classification Score

After obtaining C clusters, whose binary data are mutually exclusive in class m, the probabilistic model for classification is defined as follows:

$$\Pr(c, m|\mathbf{a}) = \Pr(m|\mathbf{a}) \Pr(c|m, \mathbf{a}). \tag{41}$$

Eq. (41) is the probability that the observation $\mathbf{a}$ belongs to class m and cluster c. The probabilistic model breaks into two parts: 1) the probability of class m given $\mathbf{a}$ modeled by a categorical distribution and 2) the probability of cluster c given m and $\mathbf{a}$ modeled by a Bernoulli distribution. The cluster c is a subset of class m. Each part is defined as follows:

$$\Pr(m|\mathbf{a}) = \prod_{k=1}^{K} (\mathbf{w}_m^+(k))^{(\mathbf{a}(k))} (\mathbf{w}_m^-(k))^{(1-\mathbf{a}(k))}, \tag{42}$$

and

$$\Pr(c|m, \mathbf{a}) = \prod_{k=1}^{K} (\mathbf{b}_m^c(k))^{(\mathbf{a}(k))} (1 - \mathbf{b}_m^c(k))^{(1-\mathbf{a}(k))}. \tag{43}$$

Let $\mathbf{g}_m^c$ be a set of the binary column vectors, $\mathbf{a}_m^n$, which belong to cluster c in class m. The maximum likelihood estimation of the Bernoulli model for cluster c is then defined as follows:

$$\mathbf{b}_m^c = \frac{\sum_{n \in \mathbf{g_m^c}} \mathbf{a}_m^n}{N_m^c}, \tag{44}$$

where N_m^c is the number of data in cluster c. The score for class m given a single frame (column vector) $\mathbf{a}$ is found by picking the highest probabilistic score when Eq. (41) is evaluated over C clusters, and for convenience logarithmic score is used as follows:

$$l_m^*(\mathbf{a}) = \max[l_m^1, ..., l_m^C], \tag{45}$$

where

$$\begin{aligned} l_m^c(\mathbf{a}) = \sum_{k=1}^{K} \mathbf{a}(k) \log(\mathbf{w}_m^+(k)) + (1 - \mathbf{a}(k)) \log(\mathbf{w}_m^-(k)) \\ + \mathbf{a}(k) \log(\mathbf{b}_m^c(k)) + (1 - \mathbf{a}(k)) \log(1 - \mathbf{b}_m^c(k)). \end{aligned} \tag{46}$$

Finally, for a series of test vectors $\mathbf{A}_t$ (e.g., frames in an utterance), the score is calculated by taking the summation of the highest log-scores from individual frames as follows:

$$L_m^*(\mathbf{A}_t) = \sum_{n=1}^{N_t} l_m^*(\mathbf{a}_t^n). \tag{47}$$

4.2 Experiments and Results

For the experiments, the Geneva Multimodal Emotion Portrayals (GEMEP) database was used to classify 12 categories of emotions, as well as the two binary dimensions: activation and valence [44]. The hypothesis is that an emotion classifier can improve its performance when speech is analyzed at different timescales with fusion before a final classification decision. The proposed method was examined by performing three experiments. First, the proposed algorithm was examined by comparing its classification results with a Bayesian classifier using a Gaussian mixture model trained at two separate temporal analysis lengths. Second, the proposed algorithm was evaluated by merging the binary matrices resulting from multitemporal analyses. Third, the fusion of a 12-way classifier with activation and valence classifiers were performed to further explore the proposed algorithm.

4.2.1 Experiment I: Classification without Fusion

For the first experiment, classification of 12 emotional categories and binary classification for activation and valence were performed at 1) the phrase level and 2) the 400-ms level.

The purpose of the experiment is twofold: First, to compare the proposed method with a Bayesian classifier. Second, to provide baseline results to be compared with the results after fusion.

Since the GEMEP database provides emotional speech segments and their labels at the phrase level, the choice of this level for classification was made. Since there were no phonetic or word timestamps, a 400-ms analysis length with a 200-ms frame interval was chosen. The average rate of English speakers is roughly 150 wpm, and the average word duration is 400 ms [104].

In the spectral clustering stage of the proposed method, the binary column vectors ($\mathbf{a}_{m,p}$) before fusion were grouped into three clusters for each emotional class and affective dimension. By performing 3-fold cross-validation on the training set, the optimal number of Gaussian components for both the Bayesian classifier and the proposed method were found as shown in Table 15. For 400 ms-level classification, the final phrase-level classification decision was made by choosing the class label with the most votes. The classification and detection results are shown in Table 15.

Table 15: Classification and detection results on a disjoint set using a Bayesian classifier (GMM) and the proposed method (BinF) at two temporal analysis lengths before fusion.

	12 Categories		Activation		Valence	
	UWA (%)	# mix.	UWA (%)	# mix.	UWA (%)	# mix.
GMM (400 ms)	36.7	96	76.7	16	76.0	16
BinF (400 ms)	**37.6**	72	**78.2**	12	**76.9**	12
GMM (phrase)	34.2	96	73.2	16	72.1	16
BinF (phrase)	**37.2**	36	**77.7**	12	**75.0**	12

The proposed method was originally designed for fusion, and its strength is revealed when feature representations are fused before classification. Fortunately, the results show that the proposed method still outperforms a Bayesian classifier, even without fusion. It

is also interesting to note that the classifier/detector with a 400 ms-analysis length outperforms the phrase-level system in all cases.

4.2.2 Experiment II: Classification with Fusion

For the second experiment, the binary column vectors ($\mathbf{a}_{m,p}$) resulting from the two temporal analyses, were fused to form $\mathbf{A}_m$. The vectors of the phrase-level analysis were upsampled by a factor so that they have the same number of instances with the 400 ms-level analysis. The visualization of the fused matrix $\mathbf{A}_m$ for a discrete-emotion category, *interest*, is depicted in Figure 16.

For 12-way classification, 72 Gaussian components were used for the 400 ms-level analysis, and 36 Gaussian components were used for the phrase-level analysis. After fusion, the total number of Gaussian clusters was 108. In the spectral clustering stage, the binary column vectors, $\mathbf{a}_m$, were grouped into three clusters for each emotional class.

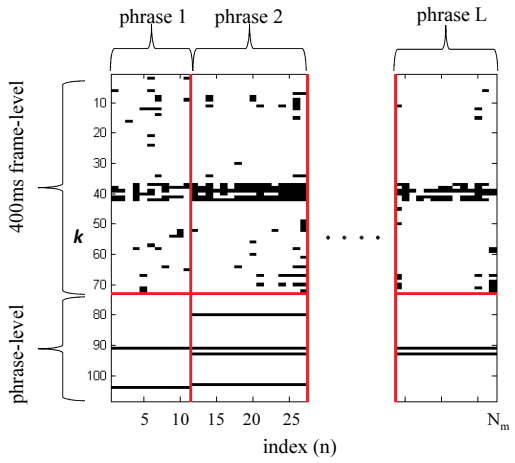


Figure 16: An example of the fused binary matrix $\mathbf{A}_m$ for an emotional state, interest.

For the activation and valence detectors, 12 Gaussian components were used for both the 400 ms and phrase-level analyses. After the fusion, the total number of Gaussian clusters was 24 for each activation and valence detector. The models were trained in the same manner as in the 12-way classifier with fusion. The results of the proposed method with fusion are shown in Table 16.

Table 16: Classification and detection results in unweighted accuracy (UWA) using the proposed method with fusion and percentage points of improvement by fusion.

	12 Categories	Activation	Valence
UWA (%) (400 ms + phrase)	44.9	83.7	80.7
# mix.	108 (72+36)	24 (12+12)	24 (12+12)
abs. improv (%)	7.3	5.5	3.8

As hypothesized, significant improvements were observed when speech was analyzed at the different timescales, along with fusion. By comparing the best results in the unweighted average accuracy obtained in the first experiment, the fusion of the two timescale analyses improved the accuracy by 7.3 percentage points for a 12-way classifier. For activation and valence detectors, 5.5 and 3.8 percentage point improvements, respectively, were observed.

4.2.3 Experiment III: Fusion of Fusion

Since there is a high correlation between emotional categories and affective dimensions [44, 105], another experiment was designed to explore the fusion of a 12-way classifier, activation detector, and valence detector. As in the second experiment, the 12-way classifier consists of 108 Gaussian clusters, and each valence and activation detector consists of 24 Gaussian clusters. After the grand fusion, the total number of Gaussian clusters was 156, and these clusters were in six different feature spaces (two temporal analysis lengths by three systems). A 12-way classifier and two detectors were trained with the binary matrices resulting from the fusion, and the results are shown in Table 17.

Table 17: Classification results in unweighted accuracy (UWA) after grand fusion and percentage points of improvement by grand fusion.

	12 Categories	Activation	Valence
UWA (%) Grand Fusion	46.1	84.2	80.7
# mix.	156	156	156
abs. improv (%)	1.2	0.5	0

Although the results do not show significant improvement, it is still interesting to note

that the proposed method does not degrade the results after a massive fusion.

4.3 Conclusion

In summary, previous studies show that different emotion-related cues are best observed at different temporal analysis lengths. Since emotional characteristics should not all be modeled at a fixed analysis frame size, a multitemporal approach to emotion classification was introduced in this chapter.

The hypothesis is that an emotion classifier or detector can improve its performance when speech is analyzed at different timescales with fusion before a final classification decision. A novel fusion algorithm, whose inputs are represented as multi-dimensional binary sequences resulting from cluster detection, was introduced and evaluated. For classifying 12 categories of emotion, the unweighted accuracy was improved by 7.3 percentage points when compared to a system with a fixed analysis frame size. For activation and valence detectors, 5.5 and 3.8 percentage point improvements, respectively, were observed. For tests of significance, 10-fold cross validation was performed using the proposed method and a Bayesian classifier. For activation and valence, p-value was less than 0.05, and for 12 categories, p-value was less than 0.025. Statistically significant improvements on all three systems validate the hypothesis, and it can be concluded that the reported fusion algorithm successfully models the multitemporal nature of emotion. Only the speech modality was explored in this chapter, but the reported fusion method is still expected to be successful for multimodal approach to emotion classification. In the following chapter, the multimodal-multitemporal approach to emotion classification will be investigated.

CHAPTER 5

MULTIMODAL-MULTITEMPORAL EMOTION CLASSIFICATION

Emotion classification is a very challenging problem because emotions are constructs with fuzzy boundaries in labels, and emotional states do not have explicit temporal boundaries [106, 107]. Moreover, emotion is expressed and perceived through multiple modalities including speech and non-speech vocalizations, gestures, facial expressions, physiological signals, and many others. Despite the challenges, recent studies show promising results in automatic emotion classification. To model the dynamic temporal process of emotion, Nwe et al. [42] employed a fully-connected hidden Markov Model (HMM). To model the slow-varying nature of emotional states, Metallinou et al. proposed a context-sensitive learning method based on HMMs and neural network algorithms [1].

The current work is mainly concerned with modeling the multitemporal characteristics of emotion using multiple modalities. The multitemporal work in speech was discussed in the previous chapter. As discussed earlier, different emotional characteristics are observed from different modalities at different timescales [95]. Physiological signals, such as the ECG, EMG, and GSR signals, contain useful emotional information when analyzed with 20-second windows [108]. In general, rapid facial expressions can be recognized from 40-ms frames, but to understand the encoded emotional state, a comparison over time is required [109].

Relatively few efforts have been reported on implementing emotion classifiers using multimodal-multitemporal information. One of the challenges with this approach is the asynchrony of feature representations which makes feature-level fusion algorithms difficult; it is even more difficult and unsuitable with high dimensional data. Late fusion algorithms, which combine the decisions of multiple classifiers, are more suitable in terms of complexity, but most late fusion algorithms require the decisions of the classifiers to be

time synchronized with the the same frame rates.

Because of the uniqueness of multimodal analysis, the fusion algorithm introduced in Chapter 4 was once again used with some modifications in this chapter. Since different characteristics of emotions are embedded in different modalities, emotion classifiers can improve their accuracy when the inputs are analyzed with different modalities at differing timescales. The objective of this chapter is to examine the effect of multitemporal analyses in both the unimodal and multimodal emotion classifiers and then to compare the reported fusion method with other fusion algorithms.

5.1 IEMOCAP Database

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database was chosen for its rich audiovisual emotion examples [1]. The database was collected with two different approaches. The first used scripted plays, where two subjects were asked to memorize and rehearse. Although scripted, the emotional content of each utterance is not pre-defined, and it depends on the interpretation of the subjects and the course of their interaction.

In the second, the subjects were asked to improvise hypothetical scenarios so that more natural emotions could be elicited. In this approach, the subjects were free to use their own words to express themselves based on their past experiments in various situations. The database was collected from 10 subjects (five males and five females), and two subjects form a pair for dyadic conversions. Each pair performed about 30 recording sessions which last about five minutes each. The five conversation pairs performed 71 scripted sessions and 80 spontaneous sessions in total. The total duration of recorded sessions is about 12 hours, and the audio sampling rate of the corpus is 16 kHz.

Along with the speech recordings, one of the subjects in the dyad wore 53 face Motion Capture (MoCap) markers according to the feature points defined in the MPEG-4 standard. The trajectories of the facial markers were captured by 8 VICON cameras at 120 frames per second. The trajectories of the markers were reconstructed in X,Y, and Z axes using the

VICON iQ 2.5 software.

The dialogues were segmented at the turn level. In total the database contains 10039 turns with an average duration of 4.5 seconds, and the average number of words per turn is 11.4. Loosely speaking, the turn-level segmentation can be also viewed as the utterance-level segmentation, where the speaker utters a thought or idea. The average duration of words in the database is about 400 ms; this gives the average speaking rate of the subjects 150 words-per-minute, which is also the average rate for English speakers in general [104].

Since the goal is to analyze and classify emotions using both audio and visual information, only the data from the speaker who wore the face MoCap markers during the dyadic sessions was used. The number of sessions is still 151, but the duration of data is halved to six hours with five thousand turns approximately.

The turn-level segments of audiovisual data were annotated with two different approaches, namely categorical and dimensional annotations. Three human evaluators annotated categorical emotions as neutral state, happiness, sadness, anger, surprise, fear, disgust, frustration, and excitement. Dimensions of valence, activation, and dominance were scaled from 1 to 5 by three human evaluators. The authors of the database employed the self-assessment manikin (SAM) to evaluate the corpus in emotional dimensions. The emotional dimensions were evaluated from 1 (negative) to 5 (positive) for valence, 1 (low) to 5 (high) for activation, and 1 (weak) to 5 (strong) for dominance.

5.2 Feature Extraction

5.2.1 Speech Feature Extraction

The acoustic features were extracted using the open-source audio feature extractor, openSMILE, developed by Technical University of Munich [43]. In addition, speech formant-related features were extracted using the formant-tracking algorithm developed at Georgia Tech [76]. Prior to feature extraction, the speech signals were pre-emphasized with a first order filter using $\alpha = 0.97$ to boost the energy of the high frequency components. After

pre-emphasis, the audio low-level descriptors (LLDs) were extracted using 30-ms Hamming windows with 10-ms overlap. The openSMILE toolkit provided energy (loudness), filter-bank energy, spectral, cepstral, and voicing related low-level descriptors every 10 ms. The trajectories of the LLDs and their first derivatives (rate of change in time) were characterized by 61 statistical and regression measures for the chunk- and utterance-level analyses as described in [45, 44].

In previous work [7], the inclusion of formant-related features resulted in an improvement in the average unweighted accuracy for classifying two levels (e.g., high and low) in four affective dimensions (activation, valence, expectation, and dominance). Also in previous work, an LPC-based formant tracker was employed to extract formant frequencies, amplitudes, and bandwidths. LPC-based algorithms produce a reasonable approximation during vowel like sounds; however, LPC-based formant trackers often encounter problems with modeling nasalized phonemes and give inconsistent results for bandwidth estimation.

Holmes et al. suggest that formant frequencies be supplemented by general spectral shape information in order to make them more useful acoustic features [73]. For this reason, a new formant tracking algorithm was introduced [76]. This algorithm first estimates formant parameters using a Gaussian mixture model. The estimates are refined using a maximum a posteriori (MAP) adaptation algorithm. The resulting Gaussian mixture model is a weighted sum of K component Gaussian densities as follows:

$$g(x_t \mid \lambda) = \sum_{k=1}^{K} w_k \mathcal{N}(x_t \mid \mu_k, \sigma_k), \tag{48}$$

where formant frequencies are obtained from the means of the Gaussian mixtures, μ_k, and the amplitudes are their weights, w_k. The formant bandwidth estimates are proportional to the standard deviation, $\sqrt{\sigma_k}$.

After obtaining the first three formant frequencies, amplitudes, and bandwidths, their interrelations were described by the differences and the ratios as shown in Table 2 in Chapter 2. The 18 low-level descriptors were then characterized by 14 statistic and 4 regression measures for chunk and utterance-level analyses as shown in Table 18.

Table 18: List of statistical and regression measures applied the formant-related LLDs and the MoCap markers.

Type	Measure
Statistical measure	max value, max location*, min value, min location*, mean, standard deviation, kurtosis, skewness, flatness, 1^{st}, 2^{nd} and 3^{rd} quartiles, interquartile range, 1^{st} and 99^{th} percentiles, and RMS value
Regression measure	slope of lin. regr., lin. regr. err., 1^{st} quad. regr. coeff., and quad. regr. err.
Other*	zero-crossing rate after mean subtraction.

*: only applied to the MoCap markers

The acoustic features are grouped into 6 categories, and their numbers are shown in Table 19. Note the differences among *energy*, *spectral energy*, and *spectral* features; *energy* is the loudness of speech signal, *spectral energy* is the sum of spectrum in a pass-band region when the signal is band-pass filtered, and *spectral* features are general statistic and harmonic measures of the spectrum. Altogether, 6697 acoustic features were extracted at each analysis level.

The acoustic features were extracted with the 400-ms, 800-ms, and utterance-level analysis frames. Since the IEMOCAP database provides the emotional speech segments and their labels at the utterance level, the choice of the utterance-level analysis frame was made. In general, the average rate of English speakers is roughly 150 words-per-minute, which suggests that the average word duration of English speakers is approximately 400 ms [104]. Therefore, the 400-ms and 800-ms analysis frames can be viewed roughly as word-level and two-word-level frames.

5.2.2 MoCap Feature Extraction

The IEMCOAP corpus contains 53 facial markers in a 3-D coordinate system captured at 120 frames-per-second. The markers were normalized for head rotation and translation and the nose marker tip was defined as the local coordinate center of each frame [1]. With

Table 19: Six groups of acoustic-prosodic LLDs with the numbers of extracted features.

Group	Speech LLDs	Num
Energy (loudness)	sum of auditory spectrum sum of RASTA filtered spectrum RMS energy	300
Spectral	roll-off point 0.25, 0.50, 0.75, 0.90 flux, entropy, variance, skewness, kurtosis, slope, psychoacoustic sharpness, harmonicity	1300
MFCC	Mel-freq cep. coeffs 1-14	1400
Spectral energy	RASTA filtered energy in bands 1-26 energy in 250-650Hz and 1 k-4 kHz	2800
Voicing	F0, probability of voicing, HNR, jitter(local, delta), shimmer(local), zero-crossing rate	573
Formant	18 formant LLDs	324

the nose tip fixed at the origin of the coordinate system, 52 facial markers in 3-D provide 156 trajectories. Unfortunately, some of the markers were lost during the recording due to various factors such as sudden movements of the subjects, the locations of the cameras, and occlusions [3]. The authors of the database report that 1.9% of the rightmost eyebrow markers, 12.5% of the right eyelid markers, 12.0% of the left eyelid markers, and 1.9% of the rightmost mouth markers were lost during the recording [3].

To accommodate this problem, the missing markers were replaced with the average of 6 nearest neighbors (3 before and 3 after). Moreover, a 6-point moving averaging filter was applied to the MoCap marker trajectories for smoothing. The MoCap markers are grouped into 6 facial regions: forehead, eye, cheek, nose, mouth, and chin as shown in Fig 17.

The MoCap data were analyzed at 4 different temporal lengths: 50-ms, 400-ms, 800-ms, and utterance levels. For subtle expressions, the MoCap data were analyzed using 50-ms analysis windows with 50% overlap. Since a 50-ms analysis window only contains 6 MoCap frames, neither statistical nor regression measures were taken, but the first derivatives of 156 trajectories of the facial markers were calculated to provide 312 features in total.

For the other frame-level and the utterance-level analyses, the statistical and regression

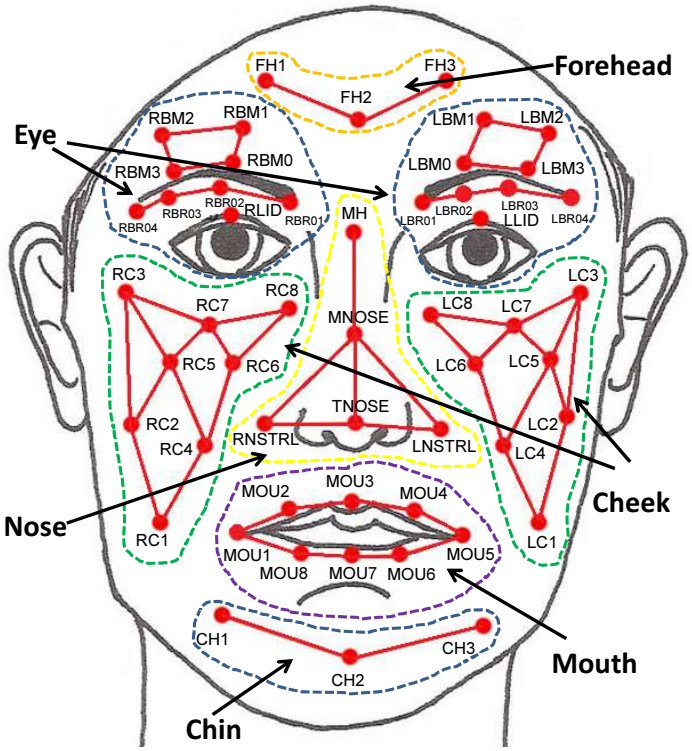


Figure 17: MoCap markers are grouped into 6 facial regions. (The figure is adapted from [3].)

measures of the MoCap trajectories and their deltas were calculated as were done for the formant-related LLDs. In addition, the locations of maximum and minimum values were found and divided by the analysis temporal length. Moreover, the mean value was subtracted from the trajectory, and the zero-crossing rate was calculated as shown in Table 18. For the MoCap data, 21 measures were calculated to provide 6552 features in total. Except of the utterance-level analysis, all of the frames were overlapped by 50%.

5.3 Unimodal-Unitemporal Classifiers

The corpus was evaluated with both the categorical and dimensional attributes as described in Section 5.1. However, approximately 17% of the data do not have a categorical label with a majority-vote agreement of the evaluators, which suggests that it is not possible to reliably classify the state of emotion with a single categorical label [1]. Furthermore, much of the research literature does not always agree on the choice of an emotion lexicon,

and such lack of agreement has produced conflicting research results and difficulties in promoting repeatable research.

The dimensional approach provides more general descriptions of an emotional expression. It has recently increased in popularity, and it is now widely accepted by many researchers. Three evaluators including the speaker him/herself assessed the corpus in three emotional dimensions with the scale from 1 to 5. Furthermore, the authors of the corpus employed the SAM system for low standard deviation and high inter-evaluator agreement. The SAM system uses a text-free assessment method, and it bypasses each evaluator's individual understanding of linguistic emotion labels [3, 110].

The authors of the corpus also analyzed cross evaluation results across the *self* and the other two evaluators by estimating the differences in reliability measures. Their results show that there are significant differences between *self* and *others* evaluations, and Cohen's kappa value decreased when the self-evaluations were included in the estimation; their results reveal a mismatch between the expression and perception of emotions [3].

In this chapter, the dimensional approach was employed, in which two evaluators (excluding *self*) labeled the utterances with the scale from 1 to 5 in the three dimensions: valence, activation, and dominance; the mean values of the two were taken afterward. As suggested in [1], the five levels of the emotional dimensions were grouped into three due to the sparsity of data in the extremes of the scale range. The first level contains ratings in the range [1, 2], the second level contains ratings in the range (2, 4), and the third level contains ratings in the range [4, 5]. The number of utterances in each level of affective dimensions is shown in Table 20.

Table 20: The distribution of the IEMOCAP database with the 3-level scale in the three emotional dimensions.

	Valence			Activation			Dominance		
Levels	names	num. of utt.	%	names	num. of utt.	%	names	num. of utt.	%
Level 1: [1, 2]	negative	1809	36%	low	566	11%	weak	516	10%
Level 2: (2, 4)	neutral	2221	44%	medium	3489	69%	neutral	3167	63%
Level 3: [4, 5]	positive	1012	20%	high	987	20%	strong	1359	27%

The three-level scale alleviates the sparsity problem, and yet most of the data belong to level 2; the amount of data in the other two levels are sufficient for model training. This is common and natural in an emotion database.

5.3.1 SVMs for Imbalanced Dataset

Support Vector Machines (SVMs) were employed to classify the three levels of the three emotional dimensions. SVMs are widely used in many disciplines for their high accuracy, ability to deal with high-dimensional data, and flexibility in modeling diverse sources of data [111, 112]. Seven individual SVM classifiers were trained depending on the modality and the analysis temporal length (3 for speech and 4 for MoCap).

As shown in Table 20, the dataset is highly imbalanced even with the three-level scale especially in the activation and dominance dimensions. Imbalanced datasets can present a challenge when training a classifier and SVMs are no exception [111]. The problem of imbalanced datasets has been approached from two main directions. The first approach is to preprocess the data by under-sampling the majority instances or over-sampling the minority instances. One of the most popular methods with this approach is Synthetic Minority Oversampling Technique (SMOTE), which creates new instances between an instance and its nearest neighbor.

The second approach is to train SVMs by assigning different misclassification costs to each class. The total misclassification cost, $C\sum_{i=1}^{l}\xi_i$, is replaced with M terms, one for each class. For M classes, the SVM solves the following optimization problem given a set of instance-label pairs $(\boldsymbol{x}_i, y_i)$ for $i = 1, ..., N$, $\boldsymbol{x}_i \in \mathbf{R}^n$, and $y_i \in 1, ..., M$ as follows [112]:

$$\begin{aligned} \min_{\boldsymbol{\omega}_m, \xi_i} \quad & \frac{1}{2}\sum_{m=1}^{M}\boldsymbol{\omega}_m^T\boldsymbol{\omega}_m + C\left(C_m\sum_{i\in\mathbf{g}_m}\xi_i\right) \\ \text{subject to} \quad & \boldsymbol{\omega}_{y_i}^T\boldsymbol{x}_i - \boldsymbol{\omega}_m^T\boldsymbol{x}_i \geq e_i^m - \xi_i, i = 1, ..., N, \end{aligned} \tag{49}$$

where N is the total number of instances, $\mathbf{g}_m$ is a set of data in class m, and

$$e_i^m = \begin{cases} 0 & \text{if } y_i = m, \\ 1 & \text{if } y_i \neq m. \end{cases} \tag{50}$$

The soft-margin constant for each class, C_m, needs to be chosen, such that the total penalty for each class is equal for imbalanced data [111, 112]. Assuming the number of misclassified examples for each class is proportional to the number of examples in each class, C_m is defined as

$$C_m = \frac{\frac{1}{N_m}}{\sum_{m=1}^{M} \frac{1}{N_m}}, \tag{51}$$

where N_m is the number of instances in class m. The global cost parameter C is found heuristically by sweeping; a smaller value of C increases the margin by ignoring data close the boundary, and C_m pushes the hyperplane further away from the minority class.

Both the SMOTE and the cost-weighting methods were evaluated with the IEMOCAP database, and they were effective with the imbalanced dataset. They improved the unweighted accuracy by approximately three percentage points; however, there was no significant difference between these two algorithms in performance. Since its computational cost is significantly less expensive than the SMOTE method, and its performance is insignificantly different, the cost-weighting method was chosen throughout the experiments in this chapter.

SVMs belong to the general category of kernel methods that employ the data only through dot products, and in the case of SVMs, the dot product can be replaced by a kernel function [111]. The choice of the kernel function is important in both the classification performance and the computational cost. Since the size of the IEMOCAP dataset is quite large both in the feature dimension and the number of instances, a linear kernel method was chosen as suggested in [112].

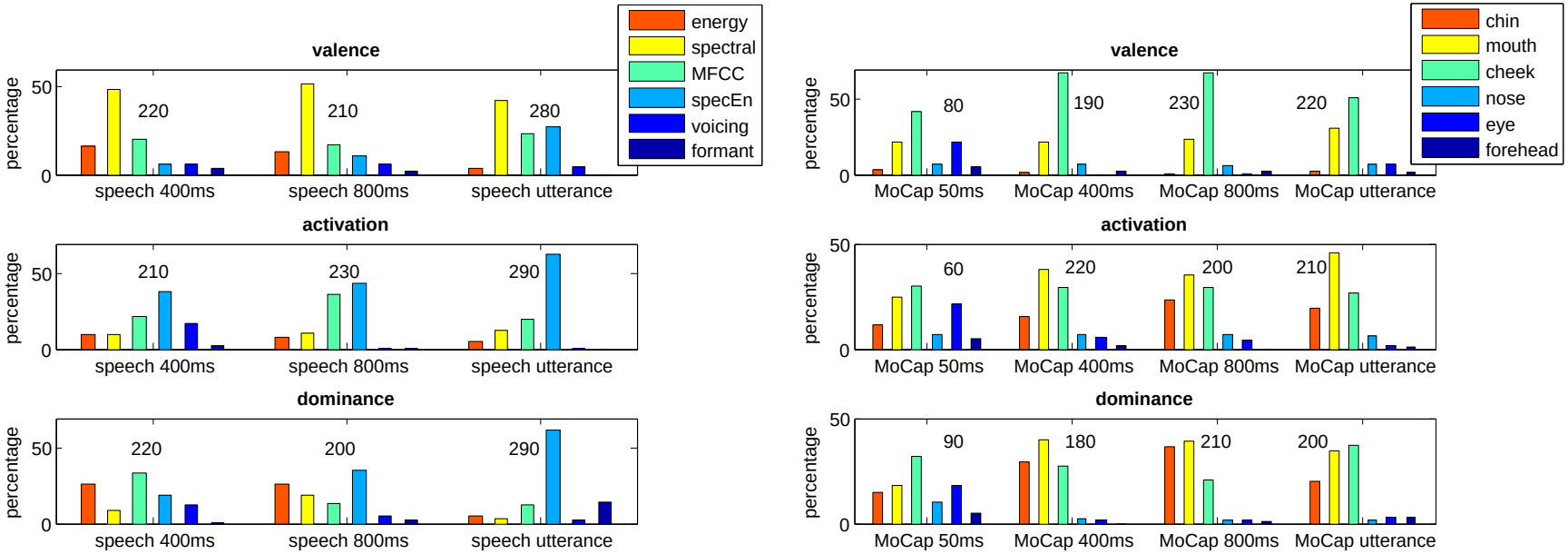
5.3.2 Feature Analysis

In general, a larger number of the features does not always result in better classification. In many cases, too large a number may cause an overfitting problem, where a training model exaggerates minor fluctuations in the data. Two main approaches exist for feature dimensionality reduction. The first approach is a group of feature projection algorithms, such as

principal components analysis (PCA) and linear discriminant analysis (LDA). The feature projection algorithms transform the data into a lower-dimension representation using linear combinations. However, this group of algorithms makes feature-selection results difficult to interpret.

A second approach uses a group of feature-ranking-based algorithms. In Chapter 2, three feature-ranking algorithms were evaluated. Two of those algorithms are the maximal-relevance method (MR) and the minimal-redundancy-maximal-relevance method (mRMR). They each select a subset of features based on mutual information gain. The other feature-ranking algorithm is based on the predictive power of features, where the features are ranked according to the unweighted accuracy when they are solitarily trained. All three algorithms were shown to be effective in feature reduction and accuracy improvement. Each algorithm obtained a different number of features in its optimal subset; however, there was not a significant difference among them in accuracy. One major advantage of feature-ranking algorithms is that their feature-selection results are interpretable. For feature interpretation and analysis purpose, the feature-ranking algorithm based on the unweighted accuracy was employed.

The SVMs were employed to rank the features by their unweighted accuracy over 10 subjects using a leave-one-out cross-validation (LOOCV) technique. This ranking process was done separately for all the 21 classifiers (3 for speech and 4 for MoCap x 3 emotion dimensions). To find the optimal subset of the ranked features, a sequential forward selection algorithm was used. Starting with the highest ranked feature, a subset of features was found by including the next highly ranked features, and the subset of features was evaluated by training a classifier. Since this wrapping processing is computationally expensive and slow, to speed up the process at each iteration, the next 10 ranked features were included in the subset for evaluation. The process was stopped when the change in the unweighted accuracy was saturated. The percentage of number of features selected in each feature group for the 21 classifiers is shown in Figure 18.



(a) Speech

(b) MoCap

Figure 18: The proportions and the total number of the selected (a) speech features analyzed at the 400-ms, 800-ms, and utterance levels, and (b) MoCap features at the 50-ms, 400-ms, 800-ms, and utterance levels in the three emotional dimensions.

5.3.2.1 Feature Analysis in Valence

For the valence dimension, a majority of the selected acoustic features are spectral-related features regardless of the temporal analysis lengths. Recall that here the spectral features are non-energy related, and they are the group of features describing the shape of spectrum with statistical and harmonic measures. The second majority feature group in valence is the group of MFCC-based features whose primary goal is also to represent the envelope of the spectrum, also known as the shape of the vocal tract. Past literature of it indicates that the spectral features play the most important role in affect classification, especially for the valence dimension [113, 94, 114]. Since in smiled speech, the mouth widens, and the lips retract resulting in a shorten vocal tract, the envelope of the spectrum significantly changes when compared to that of non-smiled speech [115].

Throughout the temporal length analyses, a majority of the selected visual features were those in the cheek region followed by ones in the mouth region. Such a result is not surprising, but rather expected, since the zygomatic major muscle, which reaches down from the cheek bone to the lip corner, controls smiling associated with positive emotional stimuli [116]. The first 10 highly ranked visual features for valence at the utterance level consist of two MoCap markers in the lower cheek region (C2 and C4) and the lip corner markers (MOU1 and MOU5). Although the cheek-region features are in the majority, the lip corner markers were ranked as the best feature (highest unweighted accuracy) at all temporal levels.

5.3.2.2 Feature Analysis in Activation

The emotional activation of the speaker is accompanied by physiological changes that affects heart rate, respiration, and articulation [117]. One sign of high activation is a tendency to raise the vocal pitch and intensity often. Banse et al. showed that the emotions with high activation values such as despair, hot anger, panic fear, and elation, have significantly higher mean energies than those of shame and sadness whose activation values are considerably

lower [118]. Similarly, the feature-ranking results emphasize the importance of the energy-related features in the activation dimension. Moreover, the results suggest that the spectral energy features are more relevant than the simple loudness features.

The physiological change such facial flushing is a good visual indicator for activation dimension. Due to the nature of MoCap data, facial flushing cannot be identified. Higher activation (arousal) is not only accompanied by physiological changes but also by higher intensity which reflects the volume and the effort of speaking [119, 118]. The change in effort of speaking leads to the change in speech and articulation rate. With the MoCap data, such a change can be observed in the lower parts of the face, especially in the mouth and the chin regions.

Throughout the temporal length analyses, a majority of the selected visual features are those in the mouth region. In the case of the valence dimension, the lip corner markers were selected as the highest ranked features. Contrarily, none of the lip-corner markers were ranked in top 10 in the activation dimension across the four temporal length analyses; instead, the lower lip markers (MOU8, MOU7, MOU6) were ranked in top 10 at the 50-ms and utterance-level analyses, and the upper lip markers (MOU2, MOU4) at the 400-ms and 800-ms analyses. Along with these lip markers, the chin markers (CH1, CH2, CH3) were ranked in top 10 at the 50-ms, 800-ms, and utterance-level analyses. The cheek region markers were still shown to be important in the activation dimension, and the number of selected features in the mouth and chin regions are noticeably higher when compared with the valence dimension.

5.3.2.3 Feature Analysis in Dominance

In a similar manner to the activation dimension, more dominant speech is characterized by a higher intensity variability, higher pitch, higher F1, lower F1 bandwidth, and higher levels of high-frequency energy [120]. Intensity reflects both the sound produced at the glottis, and the amplification and attenuation of harmonics in the vocal tract [121]. Furthermore, it is often associated with deep and forceful respiration which generates a tense and full

voice with chest register phonation. In contrast, lower dominant speech is characterized a low power, lax, and thin voice with head register phonation [117].

A majority of the selected features were the spectral-energy features when analyzed at the 800-ms and utterance levels. The MFCC-related features, which describe the spectral property using cepstral coefficients, formed the first majority group at the 400-ms analysis. One notable difference between the feature selection results of the activation and dominance dimensions was an increase in the number of the formant-related features. The formant-related features were selected as the second majority group followed by the spectral-energy group at the utterance-level analysis.

Similar to the activation dimension, the dominance dimension is also mostly characterized by the effort of speaking which is dependent on subglottal pressure, vocal fold tension, and jaw opening [122, 119]. The lower facial regions around the articulatory muscles are again shown to be more relevant than the others. The highest ranked MoCap feature was a chin-region marker for the 400-ms, 800-ms, and utterance-level analyses; it was a mouth-region marker for the 50-ms analysis.

5.3.3 Unimodal-Unitemporal Classification Results

With a subset of features as input, each classifier was trained and evaluated using a leave-one-out cross-validation (LOOCV) technique for 10 speakers. With SVMs described in Section 5.3.1, the cost parameter C was swept from 2^{-14} to 2^{14} with increment by powers of 2 as suggested in [112]. For each classifier, the decision was made at two levels: one at its own analysis length (chunk) and the other one at the utterance level. Recall that the database was originally evaluated at the utterance level by human evaluators. The utterance-level decisions of the chunk-level SVMs were made by using a majority vote method. The results are shown in Table 21.

One important result is that the speech-based classifiers were more accurate in classifying the three levels of the activation dimension than were the visual-based classifiers. Given the same analysis temporal length, the speech-based classifiers achieved 5.85 percentage

Table 21: The average UWAs ± σ of the seven individual unimodal-unitemporal classifiers for utterance-level and chunk-level classification.

	Valence		Activation		Dominance	
Temporal	utterance	chunk	utterance	chunk	utterance	chunk
Speech @ 400 ms	51.64 ± 4.02	43.96 ± 3.53	61.80 ± 4.08	51.35 ± 2.93	45.77 ± 5.06	41.01 ± 3.51
Speech @ 800 ms	51.23 ± 4.59	46.22 ± 4.28	59.68 ± 4.07	53.22 ± 3.83	46.04 ± 5.42	41.66 ± 4.75
Speech @ utt	53.41 ± 2.97	N/A	65.41 ± 2.10	N/A	52.35 ± 4.55	N/A
MoCap @ 50 ms	62.28 ± 5.05	57.91 ± 5.30	50.00 ± 5.43	45.58 ± 3.99	41.67 ± 7.74	39.56 ± 6.33
MoCap @ 400 ms	63.81 ± 4.94	59.22 ± 5.27	54.97 ± 6.41	48.83 ± 4.41	45.39 ± 6.59	41.81 ± 6.25
MoCap @ 800 ms	64.13 ± 4.84	59.94 ± 5.57	54.98 ± 4.81	50.50 ± 4.35	46.80 ± 6.57	42.86 ± 6.79
MoCap @ utt	65.56 ± 4.96	N/A	60.22 ± 3.24	N/A	48.83 ± 5.74	N/A

points higher than the visual-based classifiers in unweighted accuracy, with a p-value less than 0.001.

In contrast, the visual-based classifiers were shown to be effective for the three-level classification of valence. On average, the visual-based classifiers outperformed the speech-based classifiers with 13.52 percentage points improvement in unweighted accuracy, with a p-value less than 0.001.

The comparison results between the speech-based and visual-based classifiers in the valence and activation dimensions concur with the traditional interpretation on emotion [1, 123]. In dominance classification, no significant differences between the two were observed; however the highest classification accuracy was obtained by the speech-based classifier trained at the utterance level.

With different temporal lengths, the classifiers trained at the utterance level are shown to be more accurate than the others across all dimensions and modalities. No significant differences were observed between the 400-ms and 800-ms analyses in either modality. In the case of visual-based classification, unweighted accuracy was observed to increase when the temporal length was also increased.

5.4 Multimodal-Multitemporal Classifier Fusion

The fusion method discussed in Section 4.1 can be still used with classifiers trained on different modalities at various temporal lengths, where each classifier is trained separately to classify M classes. Each classifier's output at its decision level is converted to an $M \times L_i$

binary matrix, where L_i is the number of decisions that the i-th classifier makes before the final decision. For example, if the final decision is to be made every 2.4 seconds, the classifier trained at a 400-ms analysis window makes 6 decisions before the final classification. At each entry, the binary matrix is assigned 0 or 1, depending on the classifier's decision for each of the M classes. After obtaining all the binary matrices from the classifiers, they are merged into one matrix either by up-sampling or down-sampling the individual matrices so that they all have the equal size. For class m, the resulting binary matrix, $\mathbf{A}_m$, is again defined as in Eq. (34). The matrix $\mathbf{A}_m$ consists of N_m column vectors. Each column vector $\mathbf{a}_m^n$ has K elements which corresponds to the number of classifiers (J) multiplies by the number of classes (M) as shown in the example in Figure 19. which shows an example of the $\mathbf{A}_m$ matrix with 3 classifiers with 3 classes after merging.

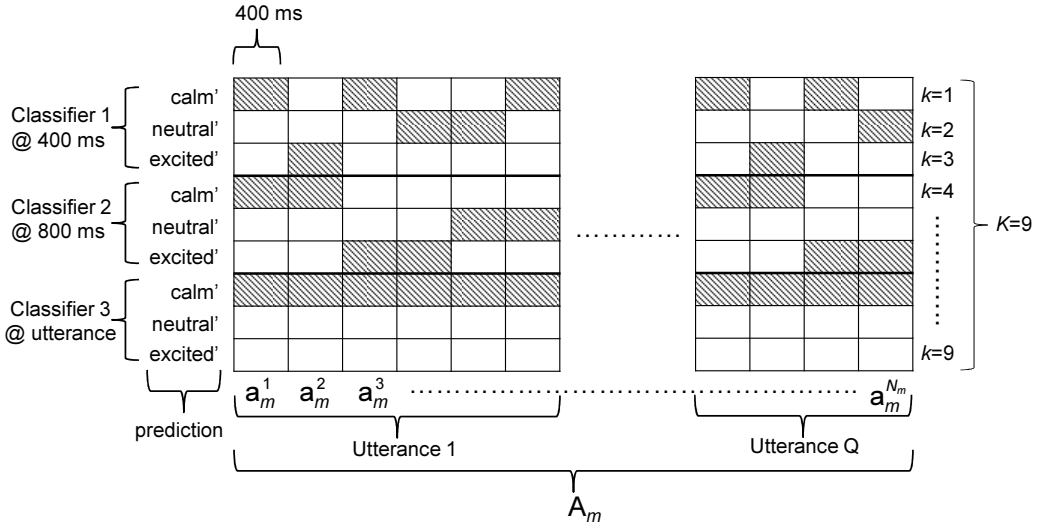


Figure 19: A graphical example of $\mathbf{A}_m$ matrix with 3 classifiers and 3 classes.

5.5 Fusion Results

5.5.1 Diversity Measures

A prevalent view on requirements for successful fusion algorithms is that individual classifiers should be *diverse*. For example, if the outputs of two classifiers are identical and in 100 percent agreement, or if they are totally opposite and in 100 percent disagreement, a fusion algorithm would not lead to success. In general, many fusion methods require the

individual classifiers to be trained on different subsets of the training data. In this work, the diversity requirement was obtained by the multitemporal analyses instead of subsampling the training data. Kuncheva et al. argue that although, many fusion algorithms assume that diversity is a key factor for success, no explicit measure of diversity is defined. Thus, they discussed the concept of diversity by comparing various methods in terms of oracle (correct/incorrect) outputs of the individual classifiers in [124].

Before presenting the fusion results, a method measuring diversity is introduced to assist in understanding the individual classifiers' effects on the fusion results. A pairwise diversity measure based on Yule's Q statistic was employed for its simplicity and for it takes into consideration both the agreement and disagreement rates between two classifiers. For the two classifiers, i and j, the diversity measure is defined as follows:

$$Q_{i,j} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \tag{52}$$

where N^{11} is the number of instances both the classifiers' outputs are correct, N^{00} is the number of instances both the classifiers' outputs are wrong, and N^{10} is the number of instances the first classifier's outputs are right while the second classifier's outputs are wrong. For two statistically independent classifiers, $Q_{i,j}$ is expected to be 0. For two classifiers tend to make more agreements than disagreements, the value of $Q_{i,j}$ will incline toward to positive, and for more disagreements, $Q_{i,j}$ will be negative. $Q_{i,j}$ is bounded between -1 and 1. The Q-values were calculated for all the possible combinations of the seven classifiers in each emotional dimension, and the average values were found over the three dimensions as presented in Table 22.

The Q-value is one way of measuring a degree of diversity between two classifiers, and they do not make any suggestions on which of the classifier pairs would lead to the highest improvement in performance. Instead, the Q-values can assist in interpreting and understanding the fusion results. The two MoCap-based classifiers trained with the 400-ms and 800-ms windows were the least *diverse* pair (with the highest Q-value) followed by the two speech-based classifiers trained with the 400-ms and 800-ms windows. This result

Table 22: The average Q-values of paired classifiers over the three emotional dimensions.

	M_{50}	M_{400}	M_{800}	M_{utt}	S_{400}	S_{800}	S_{utt}
M_{50}	1	0.79	0.73	0.47	0.18	0.17	0.20
M_{400}	0.79	1	0.89	0.56	0.34	0.32	0.25
M_{800}	0.73	0.89	1	0.61	0.35	0.39	0.29
M_{utt}	0.47	0.56	0.61	1	0.21	0.25	0.58
S_{400}	0.18	0.34	0.35	0.21	1	0.80	0.38
S_{800}	0.17	0.32	0.39	0.25	0.80	1	0.48
S_{utt}	0.20	0.25	0.29	0.58	0.38	0.48	1

M_{50}: the MoCap-based classifier trained with the 50-ms window.
S_{Utt}: the speech-based classifier trained at the utterance level.

was somewhat expected, because the classifiers in these two pairs are only varied in the temporal lengths yet closest, and the feature selection results in Figure 18 also suggest that the classifiers are closely related in terms of the proportions of the selected features. One important result is that the Q-values are lower for the pairs across the two modalities than for those within the same modality. This suggests that the fusion algorithm can *possibly* benefit more from a multimodal fusion than a unimodal fusion.

5.5.2 Unimodal-Multitemporal Fusion Results

One of the main hypotheses is that the multitemporal analyses would improve the classification performance in the emotional dimensions because different characteristics of emotions are embedded and observed at different analysis temporal lengths. To verify this hypothesis, two experiments were carried. First, to examine the effect of multitemporal approach in speech, the three speech-based classifiers were fused using the algorithm introduced in Section 4.1, and the fusion results were compared with the best unitemporal speech-based classifier which is the one trained at the utterance level. Second, the four MoCap-based classifiers were fused for the comparison with the best unitemporal MoCap-based classifier which, in this case, is the MoCap-based classifier trained at the utterance level. The results are shown in Table 23.

The unweighted accuracy was measured using a leave-one-out cross-validation

Table 23: The UWAs of the unimodal-Multitemporal fusion and the best unitemporal classifier in each modality.

	S_{utt}	S_{all}	M_{utt}	M_{all}
valence	53.41	54.36	65.55	66.27
activation	65.41	65.91	60.22	61.60
dominance	52.35	52.87	48.83	49.87
abs. improv (%)	0.70		1.04	
p-value	< 0.0025		< 0.01	

S_{utt}: the speech-based classifier trained at the utterance level.
M_{all}: the MoCap-based fusion using all the temporal length analyses.

(LOOCV) technique over the 10 subjects, and the multitemporal fusion improved the accuracy by 0.70 percentage points in the speech-based approach, and 1.04 percentage points in the MoCap-based approach. To test the statistical significance of these improvements, paired t-tests were performed. Although the improvements may look subtle, the p-values indicate that they are statistical significant; the improvements made over the 10 subjects were consistent with a small variance.

5.5.3 Multimodal-Multitemporal Fusion Results

The primary goal of this Chapter is to examine the effect of multimodal-multitemporal fusion with the proposed method. Because the performances of the individual classifiers are significantly different, and the classifiers may be redundant, fusing all the classifiers may not produce the optimal classification results. There are few studies on how to choose classifiers for fusion [125, 124]. Most of the methods are only based on statistical measures such as correlation coefficients and Yule's Q-values with an assumptions that the performance differences of classifiers are insignificant. These methods may not work properly for classifiers with significant differences in accuracy. For example, any pair of a classifier and a random number generator will produce a very high diversity value, but the fusion of those two is meaningless.

The results of the individual classifiers indicate that a certain classifier is significantly

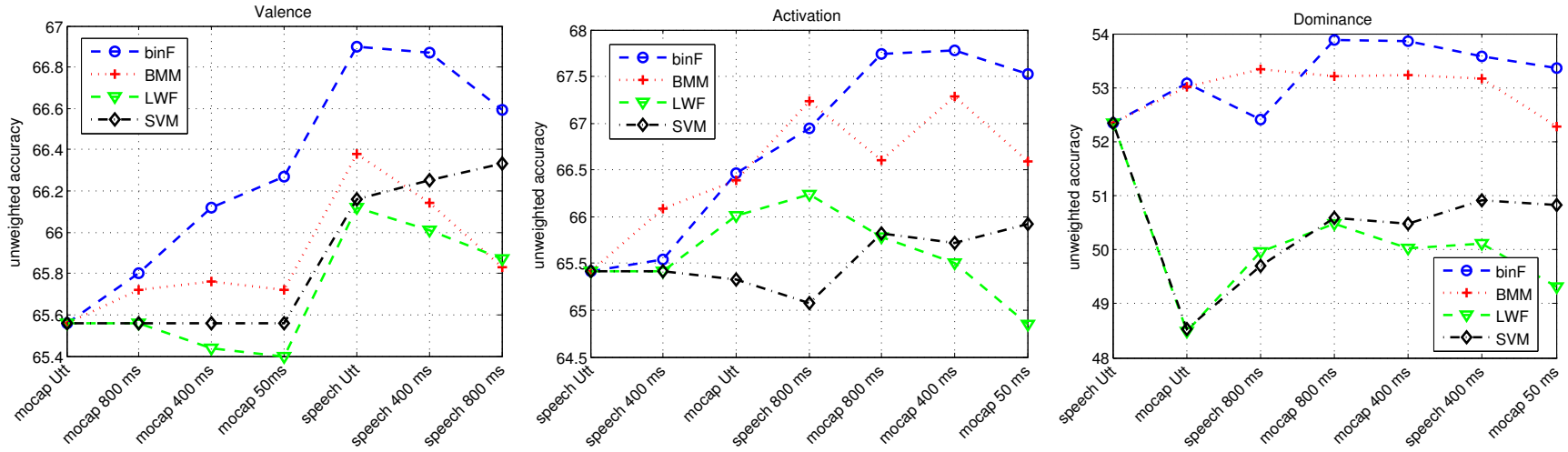
better than others; therefore, a classifier was sequentially added in the order of the unweighted accuracies. Here, the assumption is that the Q-values in Table 22 are small enough; in other words, all the classifiers are *diverse* enough.

For comparison purposes, three other methods were evaluated for fusion. The first one is a Bernoulli mixture model (BMM) on which the proposed method is partly based. Bernoulli mixture models are specifically designed for binary data classification. Since the outputs of the classifiers are converted into a binary matrix, this choice was made. For the utterance-level decisions, the chunk-level posterior probabilities of the Bernoulli mixture in each class were multiplied.

The second method is a linear weighted fusion (LWF), where weights are assigned to classifiers based on classification performance. In many studies, linear weighted fusion methods are often used to provide baseline fusion results. In the current work, the unweighted accuracy on the training set for each classifier was measured and normalized for the weights. The decisions were made using a linear weighted product of confidence measures. The third method is yet another SVM classifier. Although SVMs are generally trained with features, SVMs can be also trained with the outputs of individual classifiers for fusion.

The sequential multimodal-multitemporal fusion results are shown in Figure 20, where the order of fusion is from the classifier with the highest unweighted accuracy value to the one with the lowest value according to Table 21. The results show that all the four fusion methods benefit from the multimodal-multitemporal approach to emotion classification in the valence and activation dimensions. In the case of dominance dimension, the proposed method (binF) and the Bernoulli-mixture model (BMM) show improvements in unweighted accuracy (UWA). Each fusion method reaches its highest accuracy with a different number of fused classifiers.

For the valence dimension, the proposed method outperforms other methods and reaches 66.90% UWA using the first five individual classifiers. An interesting phenomenon



(a) Valence (b) Activation (c) Dominance

Figure 20: The sequential multimodal-multitemporal fusion results, where the order of fusion is from the classifier with the highest UWA (leftmost) to the classifier with the lowest UWA (rightmost).

is that all the four fusion methods show the highest rate of improvement when fused with the fifth classifier, which is the first speech-based classifier in the order. This phenomenon concurs with the diversity measures: the average Q-value across the modalities are lower than within the same modality.

In the case of activation, the proposed method keeps improving the classification performance until the sixth classifier fusion, where its UWA reaches 67.78%. The results of the proposed method also show that the rates of improvement are relatively high when the modalities are interlaced (e.g., speech to MoCap and MoCap to speech), and the rates are low when the modality stays the same (e.g., speech to speech and MoCap to MoCap). However, this trend is not clearly observed with the other methods.

In the case of dominance, only two methods benefit from the multimodal-multitemporal fusion. The proposed method again outperforms the others, and its highest UWA is 53.88% achieved by fusing the first four classifiers. The second best fusion method is the Bernoulli mixture model whose highest UWA is 53.59% with the fusion of the first three classifiers. Overall, the proposed fusion method outperformed the other three methods in the multimodal-multitemporal framework.

To examine the effect of multimodal-multitemporal fusion, the UWAs of the proposed method with the 10 subjects are compared to the best unimodal-unitemporal classifiers in the three emotional dimensions as shown in Table 24.

In the case of valence, the proposed method increased the UWA by 1.34 percentage points, with a p-value less than 0.025. The amount of improvement was highest in the activation dimension, where it was 2.37 percentage points, with a p-value less than 0.0025. In dominance, 1.52 percentage points improvement was observed with a p-value less than 0.05. The results show that the multimodal-multitemporal approach clearly benefits the classification task in the emotional dimensions, and the proposed fusion method is well-suited for this task.

One interesting result is that the emotion classification accuracies are noticeably higher

Table 24: The UWAs of the multimodal-multitemporal fusion using the proposed method (binF) and the best unimodal-unitemporal classifier in each dimension.

spk	sex	Valence		Activation		Dominance	
ID	(M/F)	M_{utt}	binF	S_{utt}	binF	S_{utt}	binF
1	F	65.59	69.02	69.02	71.36	48.23	52.40
2	M	63.84	67.95	65.55	64.05	54.52	53.78
3	F	65.00	67.21	68.32	69.67	48.19	53.32
4	M	55.31	55.38	64.35	66.49	46.95	46.99
5	F	66.02	66.61	62.99	66.31	58.51	58.70
6	M	70.13	70.13	63.23	66.04	48.02	48.91
7	F	68.35	69.40	65.34	69.17	59.11	59.19
8	M	67.66	69.89	66.17	68.66	50.22	49.69
9	F	73.02	71.54	66.06	67.49	53.23	57.76
10	M	60.63	61.84	63.07	68.59	56.53	58.02
Average		65.56	66.90	65.41	67.78	52.35	53.88

S_{utt}: the speech-based classifier trained at the utterance level.
M_{utt}: the MoCap-based classifier trained at the utterance level.

with female over male subjects. To examine whether the differences between two genders are significant, unpaired t-tests were performed assuming equal variances. For both the activation and dominance dimensions, the accuracies for the female subjects were, respectively, 2.0 and 4.8 percentage points higher than those for male subjects with a p-values less 0.025. In the case of variance, it was 3.7 percentage points with a p-value equals to 0.057, which is not strictly significant by traditional standards, but shows a certain trend. Much psychology and sociology literature reports that women are more emotionally expressive than men [126, 127]. The findings in the emotion classification differences between the genders do not attempt to confirm their studies on expressivity; rather the current findings are supported by them. In general, audiovisual-based emotion classifiers are designed to imitate human perception of emotion, and it is probably easier for them to classify expressive emotions than subtle ones. Similarly, the confusion matrices of the proposed method in Table 25 reflect that the classification task is relatively easier in the opposite extremes (e.g., negative vs. positive valence) than in the midrange emotions. The confusion matrices in the three emotional dimensions are shown in Table 25.

Table 25: Confusion matrices of the proposed Multimodal-multitemporal fusion method in three emotional dimensions.

Valence				Activation				Dominance			
	neg'	neu'	pos'		low'	med'	high'		weak'	neu'	strg'
neg	**61.04**	32.64	6.32	low	**76.09**	23.26	0.65	weak	**53.39**	38.43	8.19
neu	25.98	**54.70**	19.32	med	28.31	**51.35**	20.34	neu	33.01	**42.03**	24.97
pos	4.16	10.88	**84.96**	high	1.41	22.69	**75.90**	strg	13.92	19.87	**66.22**

rows: ground truth; columns: hypothesis

In Table 25, each row represents the instances in an actual class normalized by the total number of the instances, and each column represents the normalized instance in a predicted class. The results show that the opposite extremes are infrequently confused with each other, and none of the classes are overwhelmed by the other classes; the highest terms are always along the diagonal.

5.5.4 Comparison to Context-Sensitive Classifiers

To emphasize the power of the currently reported approach to emotion classification, classification results were compared to one of the most recently works utilizing the IEMOCAP database. Metallinou et al. reported their classification results on the IEMOCAP database using a context-sensitive learning method [1]. They proposed Bidirectional Long Short Term Memory (BLSTM) neural networks that take into an account an arbitrary amount of past and future audiovisual emotional expressions to recognize the current emotion of a speaker. Their rationale behind this approach was that emotional transitions are usually smooth, and emotions are slowly varying states. They reported their results on classifying the three levels of activation and valence, but not of dominance. In addition to the context-sensitive algorithm, they also implemented emotion-specific HMMs that do not make use of context information. For implementing HMMs, the authors used HTK, the popular Hidden Markov Model Toolkit. By assigning different weights to the audio and visual modalities and assuming synchronicity between them, HMMs were developed for the utterance-level classification. The results using the context-sensitive (BLSTM) and context-free (HMM) methods, along with the multimodal-multitemporal approach discussed in this chapter are

shown in Table 26.

Table 26: Audiovisual emotion classification unweighted accuracy at the utterance level using the proposed multimodal-multitemporal approach and the context-sensitive method proposed by [1].

	Valence	Activation
HMM	62.50%	60.00%
BLSTM	64.67%	52.28%
binF	66.90%	67.78%

HMM: context-free multimodal emotion classification.
BLSTM: context-sensitive multimodal emotion classification.
binF: multimodal-multitemporal emotion classification.

For valence, the proposed method improved the UWAs by 4.40 and 2.23 percentage points when compared to the context-free HMM and context-sensitive methods, respectively. For activation, the UWAs were improved by 7.78 and 15.5 percentage points when compared to the context-free HMM and context-sensitive methods, respectively. Probably due to emphasis on acoustic features in this dissertation, a higher amount of improvement was observed in activation classification. As discussed in previous sections, activation classifiers rely more on acoustic than visual features. Metallinou et al. employed a very complex system that requires multiple utterances neighboring the current utterance, but their acoustic features were not as reliable as the proposed features in this dissertation. Despite the high complexity of the context-sensitive approach, the system did not outperform the context-free approach for activation classification, whereas the proposed multimodal-multitemporal approach obtained the highest accuracies for classifying both affective dimensions.

5.6 Non-linguistic Vocalizations for Emotion Recognition

Some emotional states such as frustration, boredom, and joy seem can be identified from non-linguistic vocalizations such as sighs, yawns, laughter, and crying. Few efforts toward the automatic recognition of such non-linguistic vocalizations have recently been reported. A recent study shows that laughter can be detected with 94.4% accuracy when

tested with the Aibo Emotion Corpus of Friedrich-Alexander University (FAU-AEC) [128]. The database consists of interactions between children and Sony's pet robot Aibo [55]. Moreover, Gupta et al. [129] show that their system can distinguish emotional sighs that exist along both ends of the valence axis (positive-emotional vs. negative-emotional sighs) in the IEMOCAP database [129]. Using 26-dimensional acoustic features, which mainly consist of MFCC's, a SVM classifier was trained to classify positive-emotional sighs and negative-emotional sighs. Using leave-one-speaker-out cross-validation, the unweighted accuracy for classifying these two classes (chance = 50%) was 60.2%. Their results underscore the importance of the emotional interpretation of sighs and suggest the feasibility of using low level acoustic cues to predict the different emotional content of each sigh.

It is well-known that non-linguistic vocalizations are highly correlated with emotion; however, no effort toward human affect analysis based on vocal outbursts to our knowledge has been reported [6]. Combining a paralinguistic classifier with an emotion classifier is not an easy task, due to differences in appropriate analysis window lengths and feature sets. Since the novel fusion method discussed in this work uses weights that directly measure the outputs of classifiers in terms of the targeting class, the method would be well-suited for fusing paralinguistic and emotion classifiers.

The IEMOCAP database provides both the linguistic and non-linguistic labels with emotion. In this section, the fusion of emotion and paralinguistic classifiers is investigated in order to study the correlation between non-linguistic vocalizations and emotion.

5.6.1 Paralinguistic in the IEMOCAP Database

Nonlinguistic signals are typically associated with specific emotions, intentions, or external referents. Laughter and other nonlinguistic vocalizations are used to influence the emotional states of speakers as well as listeners, thereby also affecting their behavior [130]. The top two most frequently appearing non-linguistic vocalizations in the IEMOCAP corpus are laughter and sighs. 133 and 54 instances of laughter and sighs, respectively, were found. Prior to any further investigation, the emotional labels for these vocalizations were

found and are shown in Table 27.

Table 27: The distribution of laughter and sighs in the IEMOCAP database in three emotional dimensions.

	Valence			Activation			Dominance		
	neg	neu	pos	low	med	high	weak	neu	strg
Laughter	0	8	125	2	52	79	11	70	52
Sighs	19	18	17	11	38	5	13	39	2

As shown in Table 27, laughter is highly correlated with the level of all three affective dimensions. In valence, most of the laughter instances belong to the positive valence class while none of them are found in the class of negative valence. The result is not surprising since it is well-known that laughter is highly associated with positive valence, and when laughter accompanies an utterance, it is highly likely that the speaker's emotional state is positive. In activation, significantly fewer instances of laughter are found in the class of low activation. Laughing is related to something unexpected, and laughter comes from surprise or a recognition of an incongruity [131] which often causes one's activation level to be high. In dominance, more instances of laughter are found in the class of strong dominance than in the class of weak dominance. Laughter is a social phenomenon, and laughter rarely happens when one is alone. In contrast, people often smile when reading or even when having private thoughts [131]. Laughter, being a social phenomenon, explains why it is used to influence the emotional states of listeners [130]. Moreover, its intention is sometimes to control and dominate others by being exaggerated.

According to Table 27, sighs are somewhat correlated with the activation and dominance dimensions. Although much literature shows that people associate sighing mainly with negative, low-intensity, and deactivated emotional states [132], they do not show any correlation with the level of valence in the IEMOCAP corpus. Teigen's work [132] suggests that sighs are used to express a state of "giving up" on something or somebody; in other words, sighs are used to express low activation and a weak dominant state. The distribution of the instances of sighs is somewhat in agreement with his findings.

5.6.2 Paralinguistic Classification

As was done for emotion classification, both the speech and MoCap features were extracted using the method described in Section 5.2 for paralinguistic classification. Since the average duration of laughter and sighs in the corpus is roughly 600 ms per event, each instance was analyzed using 400-ms windows with 50% overlap. Three classes: laughter, sighs, and neither (anti-model), were trained and tested using SVMs with a radial basis function (RBF) kernel and the same cost-weighting method described in Section 5.3.1. Two classifiers were developed using speech and MoCap features separately. The classification results are shown in Table 28.

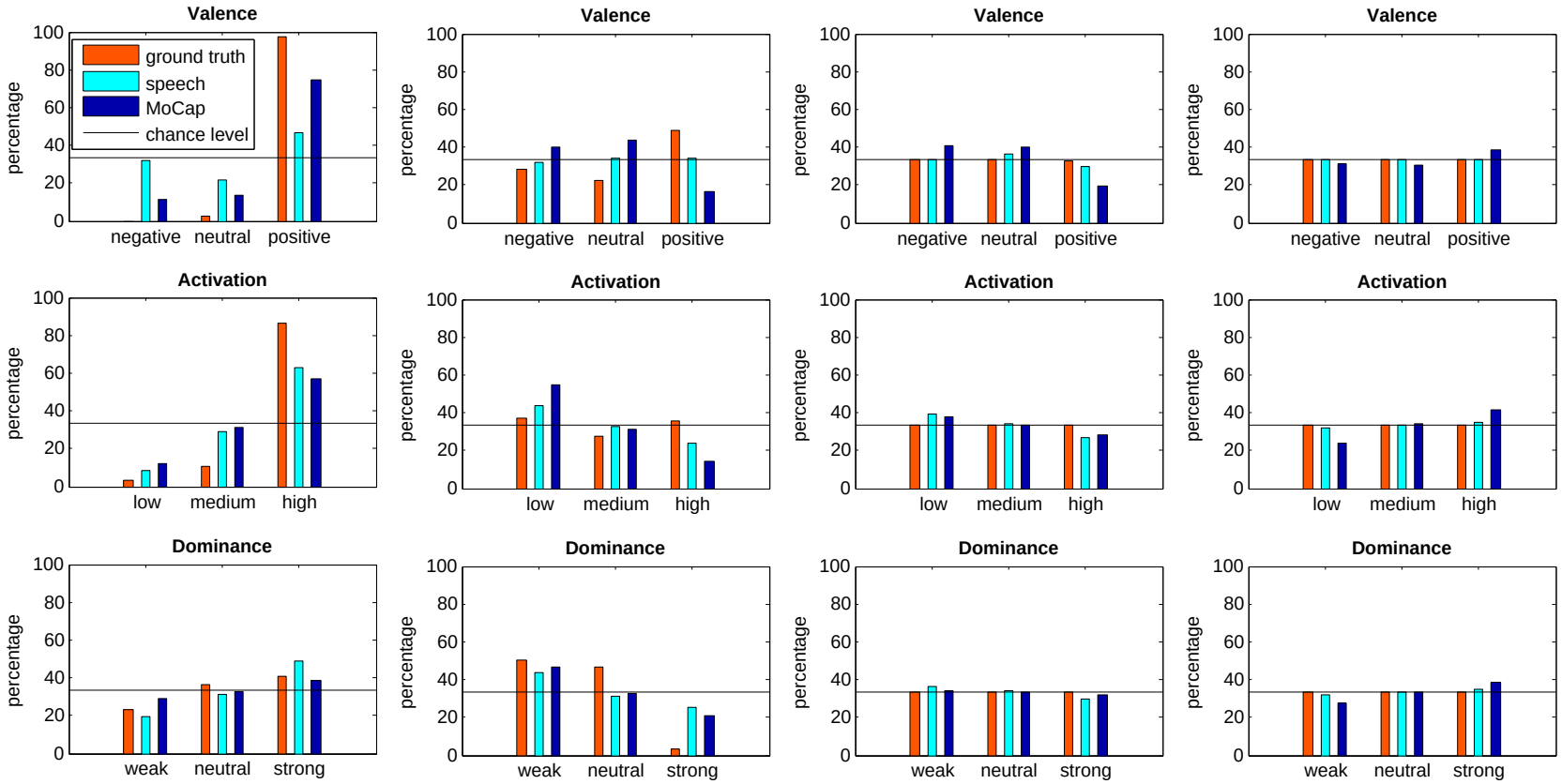
Table 28: Confusion matrices of the paralinguistic classifiers using speech and MoCap features.

	speech				MoCap		
	anti'	sighs'	laugh'		anti'	sighs'	laugh'
anti	74%	10%	16%	anti	68%	12%	20%
sighs	7%	83%	10%	sighs	11%	75%	13%
laugh	19%	6%	75%	laugh	9%	5%	87%

rows: ground truth; columns: hypothesis

Both the speech-based and MoCap-based classifiers obtained 77% unweighted accuracy; however, the confusion matrices indicate that they are different when discriminating certain classes. Since the main goal here is to provide a variety of emotional information to support emotion classification, fusion of the speech-based and MoCap-based paralinguistic classifiers was not considered.

Before fusing the paralinguistic classifiers with the emotion classifiers, the posterior probabilities of emotions given the paralinguistic classifiers' outputs Pr(emotion|paralinguistic) were found. The posterior probabilities are $\mathbf{w}_m^+(k)$ and $\mathbf{w}_m^-(k)$ as defined in Eq. (35) and Eq. (36), respectively. Recall that these posterior probabilities with equal priors are also used to weigh classifiers for fusion. If the posterior probabilities are close to the level of chance (in our case, it is 1/3), the fusion of paralinguistic classifiers with emotion classifiers will not be helpful since the outputs of the paralinguistic classifiers



(a) Pr(emotion|laughter=1) (b) Pr(emotion|sighs=1) (c) Pr(emotion|laughter=0) (d) Pr(emotion|sighs=0)

Figure 21: The posterior probabilities of affective dimensions given laughter and sighs detected by speech-based and MoCap-based paralinguistic classifiers. The results using the ground truth labels are also included. Chance would be 33.3.

convey no relevant information about the emotions in question. The resulting posterior probabilities are shown in Figure 21.

As shown in Figure 21 (a), when laughter is detected by the MoCap-based classifier, 75% of the time the associated utterance is on the positive side of the valence axis. Similarly, when laughter is detected by the speech-based classifier, 46% of the time the associated utterance belongs to the positive-valence class.

In the case of laughter, the MoCap-based paralinguistic classifier is more reliable for valence classification, whereas the speech-based classifier is slightly better for activation classification. For dominance, the detected instances of laughter do not show a strong relationship with the level of dominance; the posterior probabilities are close to the level of chance (Pr(emotion|laughter=1) $\approx$1/3).

In the case of sigh detection, both the speech and MoCap-based classifiers show trends of increases in the probability as the level of activation becomes lower and as the level of dominance becomes weaker as shown in Figure 21 (b). These trends suggest that when sighs are detected, the speaker is more likely to be in a "low-activation and weak-dominance" state. In the valence dimension, no clear trend or relationship is seen.

Based on the knowledge that neither laughter nor sighs are detected, nothing much can be inferred concerning emotional state. As shown in Figures 21 (c) and (d), the posterior probabilities for all cases are very close to the level of chance. This finding is expected since the results indicate that the presence of paralinguistic cues is helpful for classifying emotion, but their absence is less valuable.

5.6.3 Paralinguistic for Emotion Classification

To explore whether paralinguistic cues assist the emotion classification task, the best multimodal-multitemporal classification results in Section 5.5.3 were used as baselines for the valence and activation dimensions. The dominance dimension is excluded for the evaluation since no strong correlations between the emotions and ground-truth paralinguistic events (laughter and sighs) were found.

The outputs of the paralinguistic classifiers were added to the binary matrix $\mathbf{A}_m$, which consists of outputs of the multimodal-multitemporal emotion classifiers. As was done for the fusion of emotion classifiers, the binary fusion method was again used. The classification decisions were made at two levels: 400 ms and utterance. In other words, the classification decisions were made every 400 ms and for every utterance as described in Section 4.1.2. The goal of this analysis is to add additional evidence to the classifier. The classification results are shown in Table 29.

Table 29: Emotion classification results using paralinguistic cues for chunk (400 ms) and utterance-level classification.

	Valence		Activation	
	400 ms	utt	400 ms	utt
binF	61.36%	66.90%	61.67%	67.78%
binF+para	61.69%	67.03%	62.73%	67.89%
abs. improv (%)	0.33%	0.14%	1.06%	0.11%

binF: multimodal-multitemporal emotion classification.
binF+para: multimodal-multitemporal emotion classification with paralinguistic classifiers.

The results in Table 29 show that very subtle improvements were made using the paralinguistic cues. When a paired t-test was performed, no statistically significant improvement was observed for either classification level. However, for the 400 ms-level classification, a considerable trend toward significance was observed with p-value equals to 0.083 for valence and 0.070 for activation. Only subtle improvement was made because the following conditions were not fully met.

For successful fusion of the paralinguistic and emotion classifiers, three main conditions need to be met. First, large number of paralinguistic events is essential. The presence of paralinguistic cues is obviously important. As discussed in the previous section, when laughter is exhibited, it is highly likely that the subject is in a positive emotional state. However, the absence of laughter does not necessarily mean that the subject is in a negative emotional state. In the IEMOCAP corpus, only 187 out of 5,042 utterances include paralinguistic events, which is about 3.7% of the data.

Second, the duration of each paralinguistic event should be sufficiently long when compared to the duration of the classification time length. For example, if an utterance segment is two seconds long, and the associated paralinguistic event is detected only for 400 ms, the classification decision at the utterance level probably will not be impacted by the detected paralinguistic event. However, if the classification level is shorter (e.g., 400 ms), the impact of the paralinguistic detection would be greater. The classification results in Table 29 are in agreement with this notion.

Third, and most important, paralinguistics are only helpful for otherwise misclassified utterances. If the algorithm has already correctly classified the utterances associated with the paralinguistic cues, fusion with paralinguistic classifiers will not be helpful. The multimodal-multitemporal classifier (binF) misclassified only 37 utterances associated with paralinguistic events, which means that even with a perfect paralinguistic classifier, there is room for only 0.7 percentage points of improvement.

5.7 Conclusion and Discussion

To explore the effect of the proposed multitemporal approach in multimodality, the IEMOCAP database, which contains both the speech and visual information of 10 subjects was used. The emotional speech features were extracted and analyzed at the 400-ms, 800-ms, and utterance levels, and the visual (MoCap) features were analyzed at the 50-ms, 400-ms, 800-ms, and utterance levels. The features were ranked by their unweighted accuracies. In the valence dimension, the spectral features, which describe the general shape of spectrum, were highly relevant in the speech domain while the MoCap markers around the zygomatic major muscles were highly relevant in the visual domain. In both the activation and dominance dimensions, the spectral energy-related features and the lower face MoCap markers in the chin, mouth, and lower cheek regions were highly ranked.

Seven classifiers were trained separately for each unimodal-unitemporal feature set to identify the three levels of valence, activation, and dominance. For valence classification,

MoCap-based classifiers outperformed the speech-based classifiers when compared at the same temporal length. In contrast, speech-based classifiers were superior for activation. No significant differences between the two were observed in the dominance dimension. When the classifiers were compared in terms of the temporal lengths, unweighted accuracy increased as the length increased. The utterance-level classifiers were superior in all the cases.

To examine the effect of the multitemporal approach in each modality, the classifiers were fused within the same modality, and the fusion results were compared to the best unitemporal classifier. The fusion of multitemporal classifiers improved the unweighted accuracies by 0.70% points and 1.04% points in speech and vision, respectively. With the paired t-tests, the improvements were statistically significant.

The effect of the multimodal-multitemporal approach was examined by fusing the seven classifiers sequentially in each dimension. Using the proposed fusion method, the unweighted accuracies were improved by 1.34%, 2.37%, and 1.52% points in the valence, activation, and dominance dimension, respectively. T-tests show these improvements are all significant. As hypothesized, the results clearly show that emotion classification benefits from the multitemporal analyses in both the unimodal and multimodal approaches.

A novel fusion method was developed using the posterior probabilities of individual classifiers as weights. By assigning the positive and negative weights to the outputs of the classifiers, the proposed method outperformed other methods, namely Bernoulli mixture modeling and Support Vector Machines, where the accuracies of individual classifiers were not taken into account. Furthermore, the proposed method uses a spectral clustering technique due to the heterogeneous nature of data. Results show that the proposed method is superior to linear weighted fusion, which has limitations in modeling complex data. Overall, the classification results in the three emotion dimensions show that the proposed fusion method is effective and well-suited in this multimodal-multitemporal framework.

The classification results of the multimodal-multitemporal approach were also compared to the state-of-art emotion classifier proposed in [1], where the authors employed a context-sensitive emotion classifier using neural networks. For valence, the multimodal-multitemporal method outperformed the context-sensitive method by 2.24 percentage points in UWA. For activation, the multimodal-multitemporal approach improved the UWA by 15.5 percentage points.

The relationship between the diversity measures of classifiers and the rates of improvement in fusion was briefly explored. In certain cases, the rates of improvement were relatively high when the modalities were interlaced during the sequential fusion process. Although this behavior could not be fully explained by the diversity measures alone, the acquisition of more diverse classifiers would be beneficial for further improvement of fusion performance. An extension of the proposed fusion method with the classifiers trained on different subsets of data should to be considered. Furthermore, with the evidence of high diversity measures between the speech-based and MoCap-based classifiers, fusion with other modalities, such as physiological and neural signals, should be considered.

The use of two paralinguistic cues, laughter and sighs, was also investigated. The rationale behind the use of these cues is that with the knowledge of which paralinguistic cues are exhibited, emotion classification is possibly enhanced. It is well-known that laughter is highly associated with positive valence, and sighing occurs often with low activation and weak dominance; however, the fusion of paralinguistic classifiers with emotion classifiers did not impact greatly on emotion classification. Since the average duration of the paralinguistic cues are relatively short, and the paralinguistic instances are very sparse, the effect of the fusion was not much for the utterance-level classification. However, a considerable trend toward significance was observed for chunk-level classification. Paralinguistic cues are expected to be especially useful for children's emotion classification because such cues are more genuine in expressing emotion and are much more expressive than in adults.

CHAPTER 6

SOCIAL ENGAGEMENT CLASSIFICATION IN DYADIC PLAYS

Automatic emotion classification has recently received attention due to its numerous areas of application. Example applications include psychiatric diagnosis, customer relationship management, and analysis of children's behavior [11, 12]. However, only a few applications have been implemented in practice. Social and communication skills are vital in establishing social relationships needed for a healthy and productive life. Children with developmental delays, especially those with diagnoses of autism spectrum disorder (ASD), face great challenges in acquiring these skills. Research suggests that it is important to identify these conditions early on so that children can be enrolled in intervention programs, which is associated with better long term outcomes [133]. ASD does not have a clear genetic basis and can only be diagnosed based on the child's behavior. Advanced machine-learning and multimodal analyses of captured interaction sessions can enable efficient objective measurements of such behavior, potentially providing quicker diagnosis and access to treatment [134].

Many studies have shown that specific behavioral red flags early in life are associated with later diagnosis of ASD. These include decreased levels of social smiling, social initiation, orienting to name, and low eye contact [135, 136]. Lack of such behaviors can result in low levels of engagement during social interactions, and thus affect social learning. Moreover, Corbett et al. show that children with autism engaged in fewer social overtures and spent less time interacting than typically developing peers during an ecologically valid 20-minute playground paradigm [137]. Although children's engagement has been widely studied in the areas of developmental psychology and sociology, relatively few efforts have been reported on automatic engagement assessment of children's social interaction.

Recently, Guptal et al. analyzed acoustic-prosodic and spectral features from children's speech for automatic engagement prediction [138], and Hernandez et al. used a wearable

electrodermal activity (EDA) sensor to recognize ease of engagement of children during a social interaction with an adult [139]. Whitehill et al. analyzed the facial expressions of students interacting with cognitive skills training software and developed an automatic engagement classifier [140].

In this chapter, the Multimodal Dyadic Behavior (MMDB) dataset was used to automatically predict young children's level of engagement using linguistic and non-linguistic vocal cues along with visual cues, such as direction of a child's gaze or a child's gestures. The objective was to determine whether or not automatically derived measures of specific behaviors and machine learning analyses can be used to reliably reproduce subjective human ratings of children's social engagement. The relative contribution of vocal and visual modalities to predicting engagement ratings was also explored. Furthermore, the fusion of vocal and visual modalities was also performed to determine whether or not it can further improve the accuracy of engagement classification.

For engagement prediction, the vocal and visual cues of participants were represented in terms of discrete behavioral events. Novel features were extracted from these behaviors at two analysis lengths: local and stage. The local-level features indicate the co-occurrences of events, and the stage-level features indicate the durations and other statistical measures of events. For classification, a classifier fusion method introduced in Chapter 4 was employed [141]. The method was modified in the context of engagement classification with these two temporal analyses.

6.1 Multimodal Dyadic Behavior Dataset

The MMDB dataset was collected in the Child Study Lab (CSL) at Georgia Tech under a university-approved IRB protocol [142, 143]. The CSL is a child-friendly laboratory space equipped with two high resolution cameras (1920 x 1080 at 60 fps), a Kinect camera mounted on the ceiling, two omnidirectional microphones located in the center of the ceiling and at a corner of the room, and two lavalier wireless lapel microphones worn by the

child and the examiner.

The MMDB dataset consists of recordings of a brief (2-5 minute), semi-structured, table-based play interaction between young children (15-30 months of age) and an examiner. The play interaction consists of five activities (*greeting*, *ball play*, *book reading*, *hat play*, and *tickling*) designed to elicit early-emerging social communicative behaviors, such as eye contact, joint attention, smiling, vocalizations, and gestures [142].

In the greeting activity, the examiner greets the child while smiling and saying hello. In the ball activity, the examiner initiates a game of rolling a ball back and forth. In the book reading activity, the examiner brings out a book and invites the child to look through it with her. In the hat activity, the examiner places the book on her head pretending it is a hat. Lastly, in the tickling activity, the examiner engages the child in a gentle tickling game. Furthermore, in the ball, book, and tickling activities, the examiner also introduces deliberate pauses into the interaction to gauge whether and how the child re-establishes the interaction.

For each activity, the examiner assesses the child's responses by checking whether or not the child produced specific behaviors such as eye-contact, a vocal/verbal response, or a gesture. The examiner scores seventeen such behaviors as present or absent. In addition, for each of the five activities, the examiner rates the difficulty of engaging the child on a scale from 0 (easy to engage) to 2 (very difficult to engage). Early social engagement is important in the course of typical child development and is necessary for social and emotional development [144]. The MMDB dataset also includes continuous annotation of relevant child behaviors that occur during the assessment. These annotations include precise onsets and offsets of the child's gaze direction, vocalizations and verbalizations, vocal affect, and communicative gestures. These onsets and offsets are annotated by trained research assistants using the Elan [145] tool as shown in Figure 22.

To date, 121 children between the ages of 15 and 30 months have participated in this semi-structured play session, and 43 children have completed a second session 2-3 months

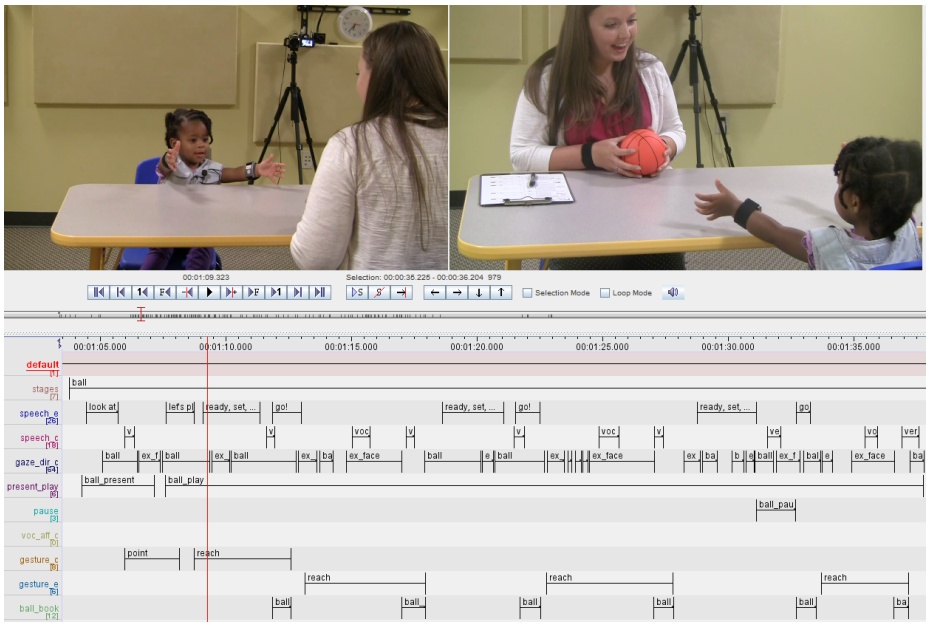


Figure 22: Screenshot of annotated Rapid-ABC session using Elan with continuous annotations of the subjects' behaviors.

later. The MMDB recordings with microphone failures, excessive visual occlusion, and extensive out-of-camera-views in video sequences were excluded from analysis. The remaining 75 sessions were used for analyzing the discrete behaviors, and features were extracted from those behaviors at local and activity levels for classifying the level of engagement. The engagement score distribution of the 75 MMDB sessions is shown in Table 30.

Table 30: Engagement score distribution in the 5 activities of 75 Rapid-ABC sessions.

	Greet	Ball	Book	Hat	Tickle
score = 0	54	66	54	73	60
score = 1	17	5	15	1	9
score = 2	4	4	6	1	6

0: easily engaged; 1: some effort was required; 2: significant effort was required.

Since the majority of the children that participated in this study were typically developing and were thus easy to engage. The distribution of engagement scores for each activity is densely populated at a score of 0. To mitigate the imbalance of the data, engagement ratings of 1 and 2 were combined into one group, which reduced the task of predicting

engagement to a binary classification (easy to engage vs. less easy to engage). Since only two examples of the hat activity fell in the less-easily-engaged group, the hat activity was excluded from the binary engagement classification.

6.2 Automatic Voice Annotation

Using automated measures of vocal (e.g., speech) and visual (e.g., eye contact, gesture) events for engagement prediction would be ideal; however, only the vocal events were automated in this work. The development of automatic visual event detection requires great effort and high computational complexity. Prior to this expensive development, it is reasonable to utilize the ground-truth annotations to examine an upper bound on accuracy of engagement classification. The work in this chapter measures the upper bound and motivates future use of automatic visual event detection. The research strategy was first to study each modality's relevance and correlation with the level of engagement in isolation, and then to assess how engagement classification would improve from a multimodal analysis.

Unimodal engagement classification was first done using the ground-truth continuous annotations produced by human observers, who were trained to reliability in behavior coding. Second, voice-related behaviors, such as onsets and offsets of examiner's and child's speech, and child's vocal affect (laughing and crying/fussing) were automatically estimated using voice activity and paralinguistic detectors. These automatically estimated voice-related annotations were then used for engagement level classification, and the results were compared to the classification results using ground-truth annotations. For multimodal engagement classification, the ground-truth and automated vocal annotations were separately fused with ground-truth visual annotations. The following subsections describe and evaluate automated measures of vocal events.

6.2.1 Voice Activity Detection

In analysis of large audio data sets, automatic segmentation plays an important role. A voice activity detector (VAD) was developed appropriate to the current environment. Off-the-shelf VADs are not general enough to use in all contexts and some level of custom design is needed for non-telephonic or non-studio voice. Since we are interested in finding onsets and offsets of the child's and examiner's voice, audio recordings using two lapel microphones were used for VAD development. One of the main challenges with the lapel microphones was that the recordings often contain rustling noise. Other unique noises associated with the recordings include those coming from bouncing balls, tapping the table, and clapping. Samples of these non-voice signals were collected for training the VAD.

Voice activity detection enables one to find the boundaries of voice so that accurate voice segmentation can be done. By imposing constraints on voice and silence durations, accurate voice activity detection can be performed, facilitating voice segmentation at the phrase level.

The main features for VAD consist of calculations of energy, zero-crossing rates, voiced/unvoiced rates, pitch, and MFCCs. These features were extracted using 30-ms Hamming windows with 15-ms overlap. Pitch and voiced/unvoiced rates were calculated using the method introduced in [146]. For voice and non-voice signals, Gaussian mixture models with a diagonal covariance matrix were trained using these features, and voice activity detection was performed every 15 ms. After detection, a 5-point median filter was used for smoothing. A number of studies have reported that the average pause duration varies from approximately 500 ms to 700 ms [147], which motivated us to choose a silence threshold at 500 ms. In other words, voice segments that were less than 500 ms apart, were considered as one segment. A pause duration is defined as the length of acoustic silence within the voice of one speaker.

The segmentation algorithm was tested on a separate dataset of child's voice previously recorded in the Child Study Lab, which consisted of an hour long recording of a child

interacting with an examiner (3936 seconds, or 01:05:36 of audio). The automatic segmentation algorithm produced 344 segments of predicted voice where the average duration of the segments was 2.85 seconds, or 980 seconds in total. This segmentation was compared to manual annotations for evaluation. The confusion matrix for the voice and non-voice classes is shown in Table 31.

Table 31: Confusion matrix of automatic segmentation in time (sec).

	Voice'	Non-voice'
Voice	901 sec	49 sec
Non-voice	79 sec	2907 sec

rows: ground truth; columns: hypothesis

The recall rate (true positive rate) is 94.88%, and the true negative rate is 97.35%. Since the segmentation accuracies are at a satisfactory level, the tool was used to segment other recordings in the dataset.

6.2.2 Cross-talk Detection

It should also be noted that the segmented voice may contain simultaneous voice from other sources. In general, four types of cross-talk exist while turn-taking occurs in interactions [148, 149]. Terminal overlaps occur when a speaker assumes that the other speaker has or is about to finish his/her turn and begins to speak. Continuers are a way of acknowledging or understanding what the speaker is saying. Conditional access to the turn occurs when the current speaker yields his/her turn or invites another speaker to interject in the conversation. Lastly, chordal is a non-serial occurrence of turns, such as laughter. Cross-talk is generally discarded in many applications of automatic speech analysis. Since common applications require analyses on a target speaker, these cross-talk events are often excluded or ignored. In our dyadic behavior analysis, cross-talk events are expected to be helpful since they contain information on how a child is responding or interacting with the examiner.

Classification of cross-talk into the four types would be useful for extensive analysis of child-adult interactions, such as those recorded in the MMDB. However, due to lack

sufficient examples of each of these four types in the MMDB, a binary cross-talk detector (cross-talk vs. no cross-talk) was developed using the lapel microphones worn by the child and the examiner. The detector consists of the voice activity detector on each channel, augmented by an estimate of the cross-correlation between the channels as well as energy measurements. When voice activity is detected from both channels, the cross-correlation is calculated to determine whether the source of the voice is a single speaker or two speakers. If there is cross-talk, the correlation between them will be low. When there is no cross-talk but voice activity is still detected from both of the microphones, the energy is calculated to determine who the speaker is. As shown in Table 32, the developed scheme provides a satisfactory level of performance for detecting this condition. The recall rate is 76.27%, and the true negative rate is 98.95%. The ground-truth for cross-talk was fully annotated by a human observer.

Table 32: Confusion matrix of cross-talk detection on segment level.

	Cross-talk'	No cross-talk'
Cross-talk	45	14
No cross-talk	3	282

rows: ground truth; columns: hypothesis

The automatic voice segmentation and the cross-talk detector were implemented together into a tool with a GUI using MATLAB making it accessible to non-speech experts. A representative example result is shown in Figure 23.

The top panel in Figure 23 shows the energy of channel one (child), and the second panel shows the pitch contour for the first channel. The second panel can be selected to plot other properties such as voiced/unvoiced ratio and SNR. The third panel shows the energy of the second channel (examiner). The red region indicates the segmented portion of the data, and the black region indicates where cross-talk occurs. The pink bar indicates the beginning of the segment, and the green bar indicates the end of the segment.

The tool also produces a table containing start and end times of the segments along with the cross-talk predictions. A typical output of the automatic cross-talk detection is shown

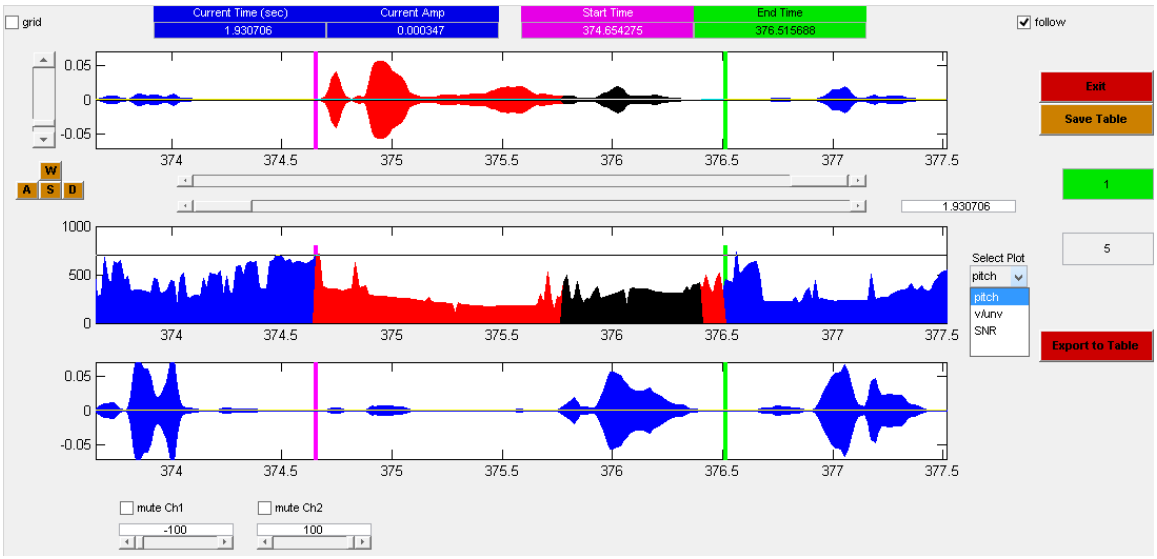


Figure 23: Graphical user interface (GUI) developed using Matlab for voice segmentation and cross-talk detection.

in Figure 24. This table can be exported to a comma separated value (CSV) file so that it can be used in other annotation tools such as Elan [145].

6.2.3 Paralinguistic Detection

829 segments of child verbalizations, 93 segments of child laughter, and 155 segments of child fussing/crying were found from the ground truth annotations of the audio streams from the 75 MMDB sessions. Tools for detecting laughter and crying in children's voices are not available in the open source community, and thus had to be developed for this study. By using the openSMILE toolkit [43], 6,373 acoustic features were extracted for detecting these paralinguistic cues. The toolkit provides energy (loudness), filter-bank energy, spectral, cepstral, and voicing related low-level descriptors (LLDs) every 10 ms. The trajectories of the LLDs and their first derivatives (rates of change in time) were characterized by 61 statistical and regression measure. The features were trained using SVMs with a polynomial kernel (p=1.5), and the laughter and crying/fussing detectors were developed using anti-models. The detection results are presented with confusion matrices as shown in Table 33.

The unweighted accuracy (UWA) for the laughter detector is 75.1%, and 74.6% for the

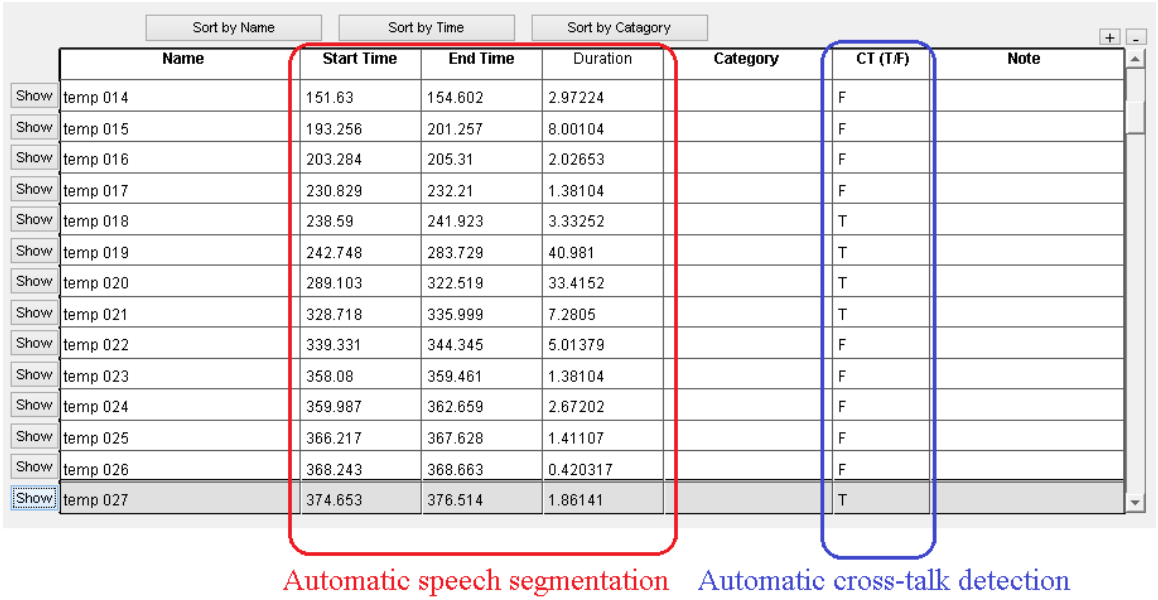


Figure 24: Voice and cross-talk detection results with timestamps, which can be exported to a CSV file.

Table 33: Confusion matrices of the laughter and crying/fussing detectors.

	anti'	laughter'		anti'	crying/fussing'
anti	79.1%	20.9%	anti	74.7%	26.3%
laughter	28.9%	71.1%	crying/fussing	24.5%	75.5%

rows: ground truth; columns: hypothesis

crying/fussing detector. The detection results were calculated using a leave-one-session-out cross-validation technique. Note that no feature selection or projection algorithms were employed for this task. Based on the findings of early related literature [128, 150], improvement is expected when an advanced feature selection algorithm is employed.

6.3 Feature Extraction

In this work, only the voice-related events were automatically annotated. For vision-related events, the ground-truth annotations were used for feature extraction. The goal of using the ground-truth annotations was to provide an upper bound on engagement classification in an ideal situation. The engagement ratings cannot be fully explained by voice and vision alone, but their contribution to engagement classification is needed to be explored for future

development.

The onsets and offsets of the child's and examiner's voice and the child's vocal affect (laughter, fussing/crying) were estimated using the voice activity and paralinguistic detectors. Using ground-truth and estimated annotations, features were extracted at two levels: local (frame) and stage. Different characteristics of discrete behavioral events can be observed at different temporal analysis lengths. The local features are the raw annotations broken down into frames that indicate co-occurrence of events (e.g., child's gaze to examiner while laughing). The stage-level features are the statistical measures of each annotation over the stage.

6.3.1 Local Binary Features

The ground-truth and estimated annotations were collapsed into 8 categories to form a binary matrix for each MMDB activity as shown in Table 34 and Figure 25. Because attending to the examiner or to the ball or the book was a good marker for actively participating in the interaction, the child's gaze direction to them was collapsed into one category, gaze-good.

Table 34: Eight local binary features from the annotations.

Modality	Features	Description
Voice	Speech-child	Child's verbalization and vocalization.
	Speech-examiner	Examiner's speech.
	Affect-pos	Child's laughter.
	Affect-neg	Child's crying and fussing.
Vision	Gaze-good	Child's gaze direction to examiner's face, hands, ball, and book.
	Obj-touch	Child's ball and book touch.
	Gesture-good	Child's reaching and pointing.
	Pause	Ball pause, book pause, and tickle pause.

The eight local binary features were trained and tested with the algorithm described in Section 4.1. The classification algorithm was initially designed with the aim of fusing multiple classifiers' outputs; however, it is also capable for training and classifying binary data such as the local binary features from the annotations.

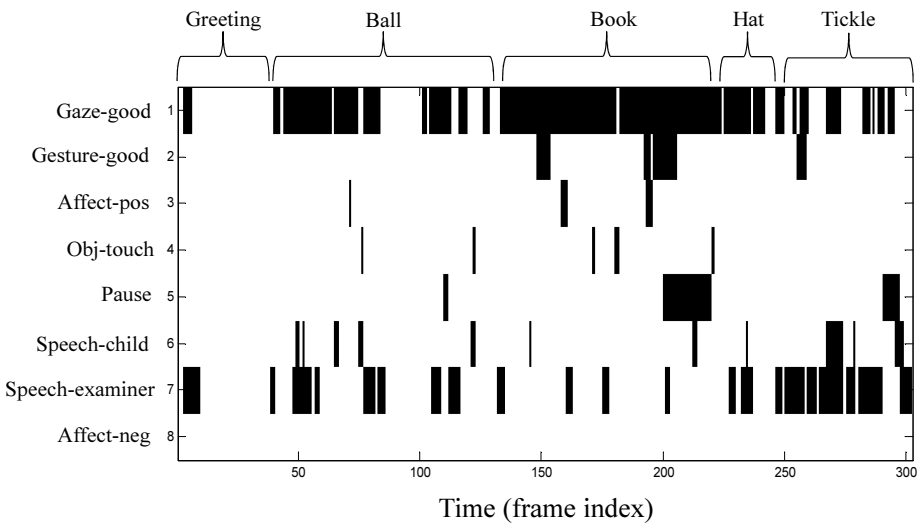


Figure 25: An example of local binary feature matrix from the annotations.

6.3.2 Stage-level Features

Since the local features can only describe the co-occurrence of events, the stage-level features were also extracted. The stage-level features are statistical measures and counts of events in each MMDB activity. These features include the measures of the child's and examiner's voice and cross-talk, child's vocal affect, child's object touch, and child's gestures which were extracted directly from the ground-truth and estimated annotations. In addition, the number of vocal turns (child to examiner and examiner to child) and number of gaze shifts (child's gaze to examiner to object, etc.) were found. These features were trained and tested using SVMs with a linear kernel. In total, 44 stage-level features were extracted as shown in Table 35.

6.4 Classifier for Engagement Classification

A classifier fusion method was introduced in Section 4.1. The method first converts classifiers' outputs into binary matrix. The method then uses the posterior probabilities of individual classifiers to find positive and negative weights. By assigning these weights to the outputs of classifiers, statistically significant improvement in accuracy are observed, as shown in Chapters 4 and 5. Since the fusion method was designed to be trained on binary

Table 35: Stage-level features from the annotations.

Modality	Num.	Features
Voice	1~5	Child' speech with 5 statistical measures.*
	6~10	Examiner's speech with 5 statistical measures.*
	11~15	Child's laughter with 5 statistical measures.*
	16~20	Child's crying and fussing with 5 statistical measures.*
	21~25	Cross-talk with 5 statistical measures.*
	26	# of vocal turns (child to examiner)
	27	# of vocal turns (examiner to child)
Vision	28~32	Child's gaze direction with 5 statistical measures.*
	33	# of child's gaze shift.
	34~38	Child's obj. touch with 5 statistical measures.*
	39~43	Child's gesture with 5 statistical measures.*
	44	Duration of activity.

*: avg. and std. of duration, max duration, # of segments, and total duration.

matrices, the same method could be used to classify the level of engagement using the local binary features. The method is briefly described below in the context of engagement classification using the local binary features. The levels of engagement to be classified are "easily engaged" and "less easily engaged," reflecting subjective engagement ratings of [0] and [1 2], respectively.

For each MMDB activity, discrete behaviors of children were annotated as either present or absent. These continuous annotations of relevant child behaviors were collapsed into eight categories as described in Section 6.3. For class m (easily engaged or less easily engaged), the resulting binary local features can be presented with a binary matrix, $\mathbf{A}_m$ as defined in Eq. (34). The size of this binary matrix is eight by N_m; N_m is the number of instances in class m. For the eight discrete behaviors, $\mathbf{w}_m^+(k)$ and $\mathbf{w}_m^-(k)$ were calculated as defined in Equations (35) and (36). The resulting $\mathbf{w}_m^+(k)$ and $\mathbf{w}_m^-(k)$ are plotted in Figure 26. These values are good indicators for measuring the relevancy of each behavior or engagement classification.

As shown in Figure 26, when a child exhibits laughter, he/she is highly like to be rated as easy to engage in the greeting and tickling activities. During the ball activity, the child touching the ball is a good indicator of ease of engagement. The child's crying/fussing is the best indicator for predicting that a child is less easy to engage in the book play. By

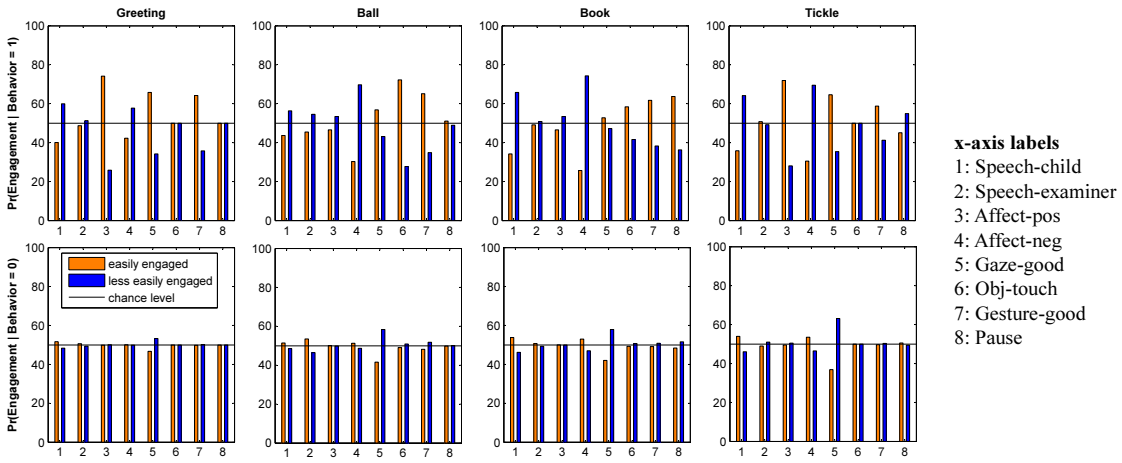


Figure 26: The posterior probabilities of engagement given each discrete behavioral event. The ground truth labels of the events are used. Chance would be 50.0.

observing the number of instances of crying and fussing in the MMDB sessions, we noted that crying and fussing came about when the examiner was putting the ball away at the end of the ball activity. The children wished to continue to play, and expressed their discontent by fussing and crying. This resulted in lower ratings of engagement for the book activity.

Often the absence of behaviors is less informative than their presence. As discussed above, when laughter occurs, it is likely that the child was rated as easy to engage in the activity. However, the absence of laughter does not necessarily mean that the child is rated as less easy to engage. In Figure 26, most posterior probabilities given the absence of behaviors are close to chance level. The exception is the child's gaze direction, whose absence is a good indicator that the child is less easy to engage. For the tickling activity, the posterior probability of less-easily-engaged given absence of gaze-good, Pr(less easily engaged | gaze-good = 0), was 63%.

6.5 Results

Prediction accuracy of the engagement levels in 75 MMDB sessions was estimated using leave-one-session-out cross-validation. The local binary features were trained and tested using the algorithm described in Section 4.1, and the stage-level features were tested using

SVM with a linear kernel. Only the voice-related annotations were automated; the vision-related features were extracted from the ground-truth annotations. For voice, the features were extracted from both the ground-truth and automated annotations for comparison. The binary classification results are shown in Table 36.

Table 36: Binary classification results in unweighted accuracy for four dyadic activities in 75 Rapid-ABC sessions using leave-one-session-out cross-validation with local and stage-level feature sets.

	VOC_{GR}		VOC_{AT}		VIS_{GR}		$VOC_{GR} + VIS_{GR}$		$VOC_{AT} + VIS_{GR}$	
	Stage	Local	Stage	Local	Stage	Local	Stage	Local	Stage	Local
GREET	64%	52%	66%	51%	72%	63%	74%	67%	73%	63%
BALL	68%	61%	72%	61%	75%	74%	76%	74%	77%	76%
BOOK	74%	72%	71%	63%	73%	66%	77%	71%	75%	71%
TICKLE	66%	70%	75%	70%	81%	79%	78%	82%	86%	81%

VOC_{GR}: features extracted from voice ground-truth annotations.
VOC_{AT}: features extracted from automatic voice and paralinguistic detectors.
VIS_{GR}: features extracted from visual ground-truth annotations.

With the classification results, five comparison studies were carried out: 1) ground-truth voice vs. automatic voice, 2) visual features vs. vocal features, 3) local features vs. stage-level features, 4) unimodality vs. multimodality, and 5) unitemporal vs. multitemporal.

First, to compare between the ground-truth voice annotations and automatic voice annotations, the binary engagement classification results were compared, that is, VOC_{GR} vs. VOC_{AT} and $VOC_{GR}+VIS_{GR}$ vs. $VOC_{AT}+VIS_{GR}$. The goal is to examine whether or not the automated voice annotations are comparable to the human annotations when the features are extracted from them for engagement classification. Automatic voice and paralinguistic detectors are not perfect in estimating onset and offset times. They also produce type I and II errors. However, their outputs are consistent in terms of making errors, and the features extracted from them turn out to be as effective as the features extracted from the manual annotations. With a paired t-test, no statistically significant differences between the two could be observed ($p > 0.05$) when the features were extracted from them for classifying engagement.

Second, to analyze which modality is more relevant to predicting children's ease of

engagement in the MMDB dataset, the classification results of VOC_{GR} and VIS_{GR} were compared to each other. Since only ground-truth annotations are available for the visual feature descriptions, ground-truth annotations for voice were also used as a fair comparison. The visual features (both at the local and stage levels) discriminate the level of engagement better than do the vocal features in the greeting, ball, and tickling activities. Since the book activity is oriented more to voice than vision, the classification results show that vocal features are more relevant in this activity. On average, the visual features produced seven percentage points higher unweighted accuracy than the vocal features with a p-value less than 0.025.

Third, the local binary features were compared to the stage-level features. The local features only describe the co-occurrence of events, whereas the stage-level features are more informative since they are characterized by more measurements and a longer period of observation. The stage-level features are statistically significantly better than local features in predicting the child's ease of engagement ($p < 0.001$). On average, stage-level features produced five percentage points higher unweighted accuracy.

Fourth, to examine the effect of the multimodal features, the best unimodal classification results were compared to the multimodal classification results, that is, $\max(VOC_{GR}, VIS_{GR})$ vs. $VOC_{GR} + VIS_{GR}$ and $\max(VOC_{AT}, VIS_{GR})$ vs. $VOC_{AT} + VIS_{GR}$. The local and stage-level classification results were compared separately to those of the multimodal approach, with no classifier fusion algorithm yet applied to this task. Instead, the multimodal classifications were done in a manner of early fusion where the subsets of features were combined before classification. For both the local and stage-level features, the multimodal approaches improved the classification unweighted accuracy by two percentage points with a p-value less than 0.0025. The MMDB dataset was designed to elicit both the vocal/verbal and visual responses from the child during the dyadic activities, and the multimodal classification results indicate that it is crucial to use both the modalities for analyzing the children's engagement in this scenario.

Lastly, the outputs of the classifiers trained with the stage-level features were fused with local binary features. Since the local binary features are not the outputs of the classifiers, their weights in Eq. (35) and Eq. (36) are considerably low while the outputs of the classifiers trained with the stage-level features are high. If the outputs of the stage-level classifier are directly fused with the local binary features, the fusion may not lead to success. This is likely because the outputs of the stage-level classifier dominate the decisions, and the local binary features have no impact on fusion results. To prevent such domination, the stage-level classifiers were trained separately with the 12 cues defined in Table 35. The resulting binary matrix for fusion is shown in Figure 27 and the classification results are shown in Table 37.

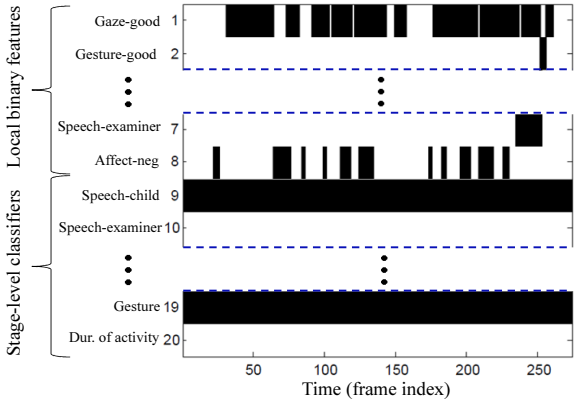


Figure 27: Multitemporal fusion using the local binary features and the classification outputs of stage-level classifiers.

Table 37: Temporal fusion results in unweighted accuracy for four dyadic activities in 75 Rapid-ABC sessions using leave-one-session-out cross-validation.

	$VOC_{GR} + VIS_{GR}$			$VOC_{AT} + VIS_{GR}$		
	Stage	Local	Stage+Local	Stage	Local	Stage+Local
GREET	74%	67%	76%	73%	63%	74%
BALL	76%	74%	78%	77%	76%	80%
BOOK	77%	71%	78%	75%	71%	76%
TICKLE	78%	82%	83%	86%	81%	85%

VOC_{GR}: features extracted from vocal ground-truth annotations.
VOC_{AT}: features extracted from automatic voice and paralinguistic detectors.
VIS_{GR}: features extracted from visual ground-truth annotations.

The temporal fusion improved the classification accuracy (UWA) in most cases except for the tickling activity with the automatic voice and paralinguistic detectors. To analyze whether the multi-temporal fusion benefit the classification, the temporal fusion results were compared to the best unitemporal results using a paired t-test. On average, the unweighted accuracy was improved by 1.4 percentage points with p-value less than 0.01.

6.6 Conclusion

The Multimodal Dyadic Behavior dataset contains semi-structured adult-child interactive behaviors. These discrete behaviors (e.g., vocalizations, gestures, eye contact, etc) were annotated with precise onsets and offsets by trained human evaluators. For each activity of the protocol, the examiner rates the effort required to engage the child using a 3-point Likert scale, with a score of 0 indicating that the child was easily engaged and a score of 2 indicating that significant effort was required. The presence and absence of the discrete behaviors are good indicators to predict how easily the child was engaged during each activity.

In this work, the voice-related behavioral events were automatically estimated using voice and paralinguistic detectors, and the features at local and stage levels were extracted. For visual feature extraction, the manually annotated ground-truth annotations were used. The results show that the features extracted from the automatic voice and paralinguistic detectors are as effective as the features from the manual annotations. It is believed that similar visual behavior detectors such as gesture, eye-contact, and object-interaction detectors can also be developed with similar effectiveness. When the voice-related features and the visual features were trained and tested separately, the visual features were superior for engagement classification in the greeting, ball, and tickling activities, whereas the vocal features were more useful in the book activity. There is no single behavioral event that contributes the most in all four activities. This result suggests that the role of modality for engagement classification is activity dependent. If a unimodal approach to engagement

classification is only allowed, the modality needs to be carefully chosen depending on the activity.

Furthermore, when both the vocal and visual features were used, the unweighted accuracy was improved by 2 percentage points from the best unimodal classifier. Although only the vocal and visual features were examined in this study, it is clear that other multimodal sensors (microphones, cameras, EDA sensors, etc) contribute different information and complement each other. Human behaviors are very complex, and cannot be explained by voice and vision alone. Thus, for improved behavior analysis, multimodal approaches must be considered.

Although the MMDB dataset contains limited and semi-structured interactive behaviors in certain activities, the range of behaviors and interactions is sufficiently rich that the techniques discussed in this chapter are expected to be generalizable to other behavior analyses. These might include clinical depression assessment, job interview analysis, and human-computer interaction evaluation.

CHAPTER 7

SUMMARY AND CONCLUSION

Emotion adds an important element to the discussion of how information is conveyed and processed by humans; indeed, it plays an important role in the contextual understanding of messages [7, 8, 9]. Whether it is the baby talk that a parent uses to communicate with a child in a higher pitch [151] or an angry customer talking to a customer service representative over the phone, affect in speech plays a key role in the development of humans. In this dissertation, qualitative and quantitative research has been carried out to discover novel acoustic features with advanced signal processing techniques and to model the multimodal and multitemporal nature of emotion with a novel machine learning algorithm. A practical application of the techniques for emotion classification was explored using social dyadic plays between a child and an adult.

In Chapter 2, the use of formant-based features for affect classification was discussed. The first tests used a LPC-based formant tracking algorithm using the SEMAINE database. A novel method of extracting formant-based features for affect classification was introduced and evaluated for binary classification in four affective dimensions. The results suggest that the formant-based features contain much information on affect.

Since LPC-based formant estimators often encounter problems with modeling nasalized phonemes and give inconsistent results for bandwidth estimation, in Section 2.3, a robust formant-tracking algorithm (GMM+MAP) was introduced to better model the formant and spectral properties. The novel formant-tracking algorithm utilizes Gaussian mixtures to estimate spectral parameters, and refines the estimates by using a MAP adaptation algorithm. The formant tracker was evaluated using the vocal tract resonance (VTR) database. When compared to a LPC-based formant tracker, the root-mean-square errors for all the phonetic classes were significantly reduced.

The reported formant-tracking algorithm was then used to extract the formant-based

features for affect classification. The classification results were compared to the LPC-based algorithm for evaluation. On average, the formant features extracted using GMM+MAP improved the unweighted accuracy by 2.1 percentage points when compared to a LPC-based algorithm. The combination of the baseline and formant features statistically significantly improved the unweighted accuracy by 2.7 percentage points, whereas the LPC-based features barely improved it by 1 percentage point. The results clearly indicate that an improved formant-tracking method will also provide improved emotion classification accuracy.

In Chapter 3, a novel method for characterizing spectral peaks was introduced. The method uses multi-resolution sinusoidal transform coding (MRSTC). Because of MRSTC's high precision in representing spectral features, including preservation of high frequency content not present in the MFCC's, additional resolving power might be present. The novel MRSTC feature set was evaluated using the GEMEP database. The classification results indicate that the MRSTC feature set alone is as effective as the baseline feature set for activation classification. The combination of the MRSTC and baseline features statistically significantly improved the unweighted accuracy by 3.5 percentage points. When the all possible combinations of the three feature sets (baseline, formant, MRSTC) were evaluated, the highest classification accuracy was obtained when all the three were combined. The results imply that these features sets are in a complementary relation, and they benefit from one another in emotion classification.

The typical spectral analysis algorithms use a fixed analysis window length, and they often fail to characterize harmonic characteristics in a broad frequency band. The multi-resolution approach enables better representation of spectral and harmonic characteristics. The experimental results clearly show that the spectral features derived from the multi-resolution spectral domains are beneficial for emotion classification.

Since emotional characteristics cannot all be modeled at a fixed analysis length, it is reasonable to analyze emotional data with multiple analysis window lengths. One of the

challenges with using multiple analysis window lengths is asynchrony of the feature representation which makes feature-level fusion algorithms difficult. Late fusion algorithms, which combine the decision of multiple classifiers, are more suitable in terms of complexity, but most late fusion algorithms require the decisions of the classifiers to be time synchronized with the same frame rates. Because of the uniqueness of the multitemporal approach in emotion classification, a novel fusion algorithm was introduced in Chapter 4.

The method converts the outputs of multitemporal classifiers into binary matrices for time synchrony, and uses weights that directly measure the outputs of classifiers in terms of the targeting class. Furthermore, the method employs a spectral clustering technique to better model the heterogeneous nature of emotional data. Using the emotional speech data in the GEMEP database, the multitemporal approach improved the unweighted accuracy by 7.3 percentage points for classifying 12 categorical emotions, and 5.5 and 3.8 percentage points improvements were observed for activation and valence, respectively. The results indicate that different emotional characteristics are embedded and observed at different analysis temporal lengths, and when these characteristics are used to train classifiers separately, the fusion of the classifiers lead to more accurate classification.

In Chapter 5, the IEMOCAP corpus was used to explore the effect of multimodal analysis for emotion classification at various temporal lengths. Seven unimodal-unitemporal classifiers (3 speech-based and 4 vision-based) were developed to identify three levels of the valence, activation, and dominance dimensions. Yule's Q-values for pairs of the classifiers were measured to understand the relationship between the diversity of classifiers and the rates of improvement in fusion. In certain cases, the rates of improvement were relatively high when the modalities were interlaced during the sequential fusion process. Although this behavior could not be fully explained by the diversity measures alone, the acquisition of more diverse classifiers would be beneficial for further improvement. Also, the classification results indicate that multimodal-multitemporal analysis is much more effective than unimodal-multitemporal analysis.

The classification results of the multimodal-multitemporal approach were also compared to the state-of-art emotion classifier proposed in [1], where the authors employed a context-sensitive emotion classifier using neural networks. For valence, the multimodal-multitemporal method outperformed the context-sensitive method by 2.24 percentage points in UWA. For activation, the multimodal-multitemporal approach improved the UWA by 15.5 percentage points.

This study has provided evidence that multimodal analysis is very useful in classifying emotion. Fusion with even more modalities, such as physiological and neural signals, should be considered if we are to better understand the expression of emotion. In addition, syntax and semantics can potentially contribute information for emotion classification. With the help of an automatic speech recognition system, semantics and syntax-based emotion classifiers could be developed. These systems would require lexicons and labeled sentences for training models of the various emotion classes. It is likely that such syntactic and semantic information would contribute to classification accuracy. Also, the introduced fusion algorithm is capable of combining classifiers trained on syntactic, semantic, voice, visual, physiological and neural signals, as well as others that may arise as instrumentation is developed.

The use of two paralinguistic cues, laughter and sighs, was also investigated. The rationale behind the use of these cues is that with the knowledge of which paralinguistic cues are exhibited, emotion classification is possibly enhanced. It is well-known fact that laughter is highly associated with positive valence, and sighing occurs often with low activation and weak dominance; however, the fusion of paralinguistic classifiers with emotion classifiers did not impact greatly on emotion classification. Since the average duration of the paralinguistic cues are relatively short, and the paralinguistic instances are very sparse, the effect of the fusion was not much for the utterance-level classification. However, a considerable trend toward significance was observed for chunk-level classification.

Paralinguistic cues are expected to be especially useful for children's emotion classification because such cues are more genuine in expressing emotion and are much more expressive than in adults. In this dissertation, only laughter and sighs were investigated. Despite their statistically insignificant improvement in classification accuracy, their use with other paralinguistic cues, such as yawns, moans, and gasps, should contribute to better results. The potential benefit of using paralinguistic cues is expected to be greater than that demonstrated in this restricted study since as the number of paralinguistic categories increases, the problem of sparsity is reduced.

In Chapter 6, the reported emotion classification methods were applied to predict engagement levels of children in dyadic plays. Using the Multimodal Dyadic Behavior (MMDB) dataset, speech-related behaviors were automatically estimated using speech and paralinguistic detectors. The speech features were extracted from the automatically estimated behaviors at the local and stage levels. The engagement classification results indicate that the features extracted from the automatic speech and paralinguistic detectors are as effective as those from the manual annotations. The visual features were only extracted from the manual annotations at this point, and it is believed that similar visual behavior detectors such as gesture, eye-contract, and object-interaction detectors can also be developed. Automatic visual behavior detectors are crucial for this engagement prediction task, since the visual features were shown to be more effective than speech in most of the activities except for the book reading. Although only the speech-related behaviors were automatically estimated, the results reveal that the same techniques explored for emotion classification are still effective for practical applications.

APPENDIX A

APPLICATION OF SPECTRAL FEATURES IN SPEECH INTELLIGIBILITY ESTIMATION

Verbal communication plays a major role in one's life style, and when it is distorted, it can also create a deficit in psychological well-being [152]. Head and neck cancer patients have problematic speech issues, such as hoarseness, dryness, excessive phlegm, coughing, or articulation difficulty [153]. Treatments for head and neck cancer vary depending on the condition of patients. For inoperable tumors of the head and neck, concomitant chemo-radiation treatment (CCRT) is an alternative treatment. Such treatment may increase speech quality over time, but it is also known that two of the early side effects of such treatment are hoarseness in speech and lack of saliva (xerostomia) [154]. Distorted speech may occur as a result of mouth dryness resulting in greater friction between the structures of the mouth, which causes the typical tone of a "thick tongue" [155]. It is useful to measure speech intelligibility for the patient both before and after CCRT.

It has been shown that acoustic features of speech, such as pitch period perturbation, amplitude perturbation, vocal noise, and spectral and formant analysis, serve as good measures for identifying pathological speech [156]. A method for extracting features in a multi-resolution sinusoidal transform coding (MRSTC) framework is presented in Chapter 3. In this appendix, the novel features are examined using the NKI CCRT Speech Corpus. These speech data are subjectively labeled, with intelligibility scores determined at the utterance level [45].

A.1 NKI CCRT Speech Corpus

The NKI CCRT Speech Corpus (NCSC) was recorded at the Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute as described in [157] and

[158]. The recordings were collected from 55 patients who underwent concomitant chemo-radiation treatment. All speakers were instructed to read a 189-word passage from a Dutch fairy tale.

Thirteen recently graduated or soon-to-graduate speech pathologists evaluated the speech intelligibility of the recordings on a 7-point scale ranging from very poor intelligibility (1) to very good intelligibility (7). The evaluators were instructed to ignore aspects of reading fluency and any interrupting noises in the recordings. For each utterance, the average of 13 evaluators' ratings was used to produce intelligibility scores ranging from 2.0 to 6.7. An Interclass Correlation Coefficient (ICC) of 0.95 for the 13 evaluators verifies that the mean score of the evaluators is reliable [158]. In total, the corpus contains 2,386 utterances. The interval-scaled intelligibility scores are available for 1,647 utterances along with the binary class labels obtained by dividing the interval-scaled scores into two classes at the median of the score distribution. The scores for the remaining 739 utterances are not fully open to the public at this point, but binary classification (*intelligible* and *unintelligible*) results can be only obtained in a blind procedure by submitting one's predictions using a defined number of trials [45].

A.2 Intelligibility Score Prediction

To evaluate the new feature set extracted using MRSTC, experiments were performed in two frameworks: one with a classifier trained and tested on binary classes and the other with the interval-scaled intelligibility scores using regression models.

The main purpose of the binary classification is to compare the prediction results using the combined MRSTC and baseline features to the prediction results presented by the organizers of the corpus. The organizers used a SVM classifier with the baseline features on the test set whose labels are hidden from the users. Thus, for a fair comparison, a SVM classifier was trained on the fully accessible data, and the predictions on the test set were submitted to the organizers to evaluate the predictions.

With the interval-scaled scores, the effects of the MRSTC features were investigated by comparing the performance of regression models before and after combining the new feature set to the baseline feature set. Furthermore, the interval-scaled ground truth scores and the predicted scores were quantized into 3 discrete classes for evaluation. All the experiments in the regression framework were performed using 10-fold cross-validation.

A.2.1 Binary intelligibility classification

To optimize classification performance, it is important to reduce the dimensionality of the feature set. The feature selection algorithm used in this section consists of three stages. First, 10-fold cross-validation was performed on each individual feature, and the average unweighted accuracy (UWA) was measured as defined in Eq. (53).

$$\text{UWA} = \frac{1}{C}\sum_{c=1}^{C}\frac{\text{\# of hits in class } c}{\text{\# of instances in class } c}, \tag{53}$$

where C is the number of classes. Each feature was ranked according to the average unweighted accuracy over the 10 folds. Second, using the sequential forward selection algorithm, a subset of features was obtained by adding features one-by-one to the subset in rank order [7, 68]. Starting with the first ranked feature, the feature x_i was added to the subset and tested for the reduction in the error. If the newly added feature failed to improve in the error rate, the feature was discarded, and the process continues to the next ranked feature for test. Finally, with the obtained feature subset from the sequential forward algorithm, a backward feature elimination was performed. In this stage, each feature was temporarily excluded from the subset and tested for the reduction in the error. If there was an improvement in the error rate, the feature was eliminated from the subset. If no elimination was necessary, the resulting subset was chosen as the optimal subset. All the error rate were tested with a 10-fold cross-validation technique, and Gaussian mixture models with 32 mixtures and diagonal covariance matrices were used for training.

To investigate how the newly developed features impact intelligibility classification, three optimal subsets were found by using 1) the baseline features alone, 2) the MRSTC

features alone, and 3) the combined features. Figure 28 shows how the three feature subsets reach to their maximum recall rates as the number of features increases.

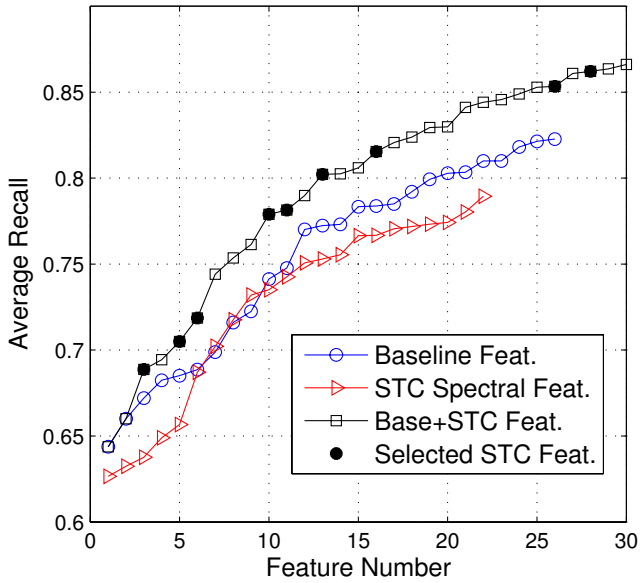


Figure 28: Feature selection results in unweighted accuracy with the number of features in each subset.

Each subset consists of different numbers of features, but the trend of each curve can be clearly seen. The reader should note that the maximum average recall rate occurs when two sets of features are combined. The optimal subsets consist of 26, 22, and 30 features of the baseline, MRSTC, and the combined features, respectively. In the optimal subset of the combined features, 9 out of the 30 features are of the newly extracted features, and 4 out of the 9 features are statistical measures of $\Delta\mathbf{f}/f_o$. This result confirms that the measurements of inter-peak distance of harmonics contain useful voice quality information. The first 10 features of the 22 selected features in the MRSTC optimal subset and the first 10 features of the 30 selected features in the optimal combined subset are shown in the Table 38.

For the combined feature set, 4 out of the top 10 features are the newly developed MRSTC features. The highest ranked MRSTC feature is of $\Delta\mathbf{f}/f_o$, the measurement of inter-peak distance of harmonics over the pitch frequency. In the combined set, 9 out of the top 10 features are related to spectral low-level descriptors (LLDs) except for the shimmer

Table 38: The first 10 features in the optimal subsets of MRSTC and combined features.

order	MRSTC	Baseline + MRSTC
1	$\Delta\mathbf{f}/f_o$ of R_{ff} in 4th FB	11th MFCC
2	log (FFT) in 4th FB	shimmer
3	log (FFT) in 3rd FB	$\Delta\mathbf{f}/f_o$ of R_{ff} in 4th FB*
4	$\Delta\mathbf{f}/f_o$ of log (FFT) in 1st FB	energy in 2nd RASTA FB
5	log (FFT) in 1st FB	$\Delta\mathbf{f}/f_o$ of R_{ff} in 1st FB*
6	$\Delta\mathbf{f}/f_o$ of R_{ff} in 2nd FB	log (FFT) in 4th FB*
7	$\Delta\mathbf{f}/f_o$ of R_{ff} in 1st FB	11th MFCC
8	R_{ff} in 3rd FB	spectral entropy
9	HIR in 3rd FB	spectral roll-off 50%
10	R_{ff} in 3rd FB	log (FFT) in 3rd FB*

FB: frequency band; statistical measures are excluded from the names.
*: selected MRSTC features in the combined feature sets.

measurement.

Two experiments were carried out to examine the effect of the proposed MRSTC features on the classification task. First, to verify that the improvement made by the inclusion of MRSTC features is significant, a paired t-test was performed using 10-fold cross-validation for results obtained by the optimal feature subset of the baseline features and that of the baseline + MRSTC features. Second, to test whether the trained model generalizes, the classification was performed on the development and test sets where the data distribution of the two classes, Intelligible (I) and Unintelligible (UI), is different from the training set. Unweighted accuracy was chosen as the main evaluation measure for the experiments because of the imbalance of data in the training, development, and test sets.

Gaussian mixture models were used for speech intelligibility classification. After multiple evaluations on the number of mixtures, 32 Gaussian mixtures were chosen with diagonal covariance matrices. The classification results of 10-fold cross-validation using the training set are shown in Table 39.

Although the average unweighted accuracy of the MRSTC feature set alone is 3.32 percentage points below that of the baseline feature set, the unweighted accuracy is improved by 4.32 percentage points when those two feature sets were combined. The p-value for

Table 39: 10-fold cross validation results on the training set using the optimal feature subsets.

	TPR (%)	TNR (%)	ACC (%)	UWA (%)
Baseline	80.2	84.3	82.6	82.3
MRSTC	80.0	78.0	78.8	79.0
Base. + MRSTC	85.4	87.8	86.8	86.6

TPR: true positive (intelligible) rate; TNR: true negative (unintelligible) rate
ACC: accuracy; UWA: unweighted accuracy

the improvement is less than 0.0025 which indicates that the improvement is statistically significant.

Using the optimal feature set found in Section A.2.1, the system was trained on the training set with 32 Gaussian mixtures modeling the 30 optimal combined features, and tested on the development set. For comparison purposes, the optimal subsets of the baseline features and the MRSTC features were also tested on the development set. The classification results on the development set are shown in Table 40.

Table 40: Classification results on the development set.

	TPR (%)	TNR (%)	ACC (%)	UWA (%)
Baseline	58.4	67.9	63.5	63.1
MRSTC	66.3	57.0	61.3	61.7
Base. + MRSTC	71.0	61.2	65.6	66.1

TPR: true positive (intelligible) rate; TNR: true negative (unintelligible) rate
ACC: accuracy; UWA: unweighted accuracy

The highest unweighted accuracy is 66.10% obtained by using the optimal combined feature set. The inclusion of the new features improved unweighted accuracy by 2.97 percentage points on the development set; the result suggests that the trained model generalizes well with a similar improvement made in the previous cross-validation test.

The top panel of Figure (29) shows the histogram of the development set with the distribution of intelligibility scores, and the bottom graph shows accuracy over each intelligibility interval. The intelligibility score range for the class I is [5.77 6.71], and for the class UI the range is [1.99 5.72].

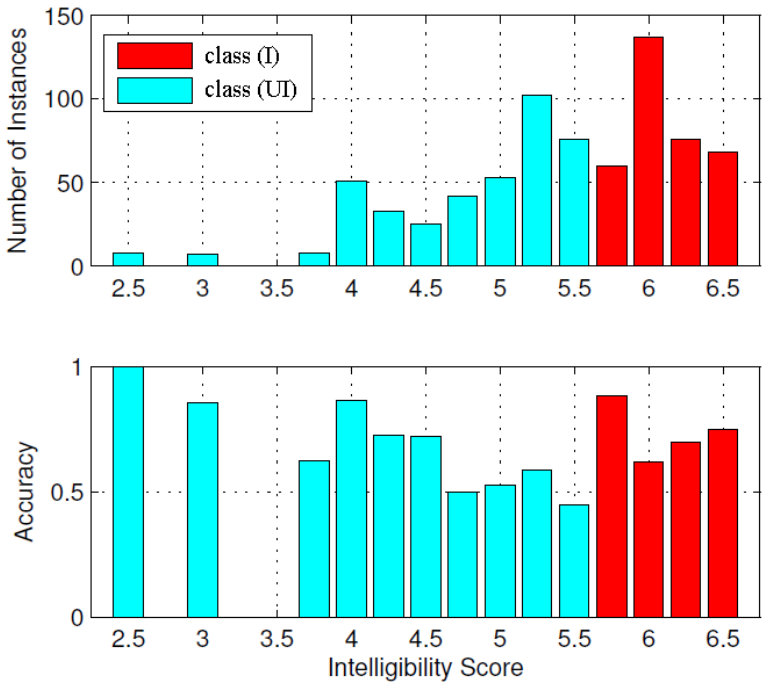


Figure 29: (top) Histogram of intelligibility scores in the development set. (bottom) Unweighted accuracy over intelligibility scores.

It is interesting to note that for the class UI, the classifier struggles the most in the score range [5.55 5.75], where it is close to the boundary between the two classes. For the class I, its accuracy rate is the highest when the score range is closest to the boundary; however, a trend of increase in accuracy as the intelligibility score moves away from the boundary can be observed. In the lowest intelligibility score range [0 2.5], where there are 8 instances, the binary classifier achieved 100% accuracy.

Ground-truth labels for the test set are not open to the public, and the binary classification results can only be obtained by submitting the predictions using a defined number of trials. Due to limited access to the data, the classification results on the test set are shown by comparing to the baseline results presented by the organizers of the corpus [45, 158]. The baseline results on the test set were obtained by using SVM and Random Forests algorithms trained on the fully available data. The main goal of this study is to evaluate the effect of the MRSTC features. Therefore an SVM with a linear kernel was trained using

the combined feature set and tested for the binary classification on the test set. No attempt was made to reproduce the Random Forests approach. The results are shown in Table 41 with the accuracy measurements, and the confusion matrix is shown in Table 42. Accuracy (ACC) is defined as (# of hits) / (# of instances), and unweighted accuracy (UWA) is defined in Eq. (53).

Table 41: Binary classification results on the test set.

classifier	SVM	Random Forests	SVM
feature set	baseline	baseline	base. + MRSTC
ACC	68.0	68.9	72.7
UWA	66.2	67.5	71.2

ACC: accuracy; UWA: unweighted accuracy

Table 42: Confusion matrix of the SVM trained with the baseline and MRSTC features.

	Intelligible'	Unintelligible'	sum
Intelligible	363	112	475
Unintelligible	90	174	264

rows: ground truth; columns: hypothesis

The results clearly show that the inclusion of MRSTC features improves the accuracy of the intelligibility classifier. Unweighted accuracy is improved by 5 percentage points when compared to the SVM trained on the baseline features, and is 3.7 percentage points higher than that of Random Forests.

A.2.2 Regression Analysis

Two regression methods were employed for intelligibility score prediction. When the number of predictors is large compared to the number of examples, a multicollinearity problem may cause the regression models to fail. One approach to coping with this potential problem is to use principal components analysis on the training data, then use the principal components to train a regression model. The approach is called principal components regression (PCR). Another well-known approach is called partial least squares (PLS) regression which finds components of predictors that are relevant to dependent variables. In

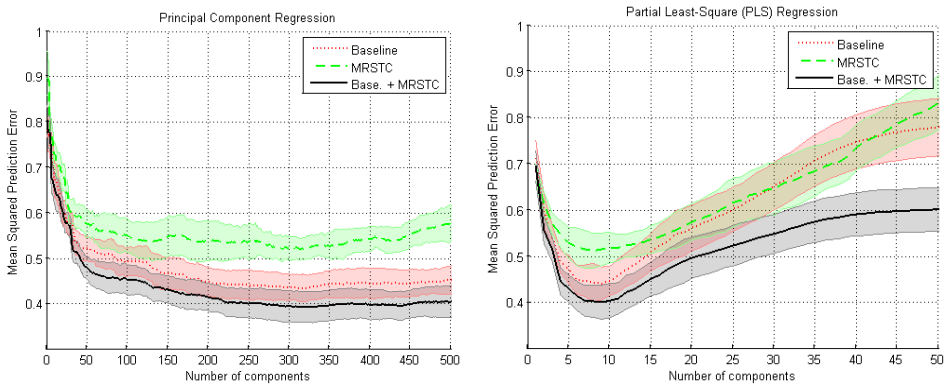


Figure 30: MSE of the (left) PCR and (right) PLS estimators with the number of predictor components. The shaded area represents the 95% confidence intervals.

general, PLS is thought to be a better regression method than PCR, since principal components are calculated using only the data without the knowledge of labels.

In order to investigate the effects of the MRSTC features, regression models were trained using both PCR and PLS with 10-fold cross validation. The regression model was obtained using the baseline features, the MRSTC features, and the combined (baseline + MRSTC) features. The mean squared estimation errors (MSE) of PCR and PLS were calculated as shown in Figure 30, where the number of predictor components increased from 1 to 500 and 1 to 50 for PCR and PLS, respectively. The shaded area represents the 95% confidence intervals of MSE.

With PCR, the lowest MSE was 0.39. It was obtained using 328 principal components from the combined feature set. The lowest MSE obtained using only the baseline feature set was 0.43. The inclusion of the new features improved the MSE by 0.04 with a p-value less than 0.001 when a paired t-test was performed. With PLS, the lowest MSE was 0.40 using 9 components of the combined feature set. This is 0.04 less than that of the baseline feature set. The p-value for the improvement is again below 0.001 indicating the improvement was highly statistically significant. Overall, regression models with the combined feature set almost always outperformed models that used only the baseline features.

When comparing the mean squared errors of the two regression methods, the difference is not significant, but PLS is much more efficient in terms of both computation and the

number of components. In Figure 31, ground-truth scores are plotted against the predicted scores of PLS using the combined feature set. The Pearson's correlation coefficient between the ground-truth and the predicted scores from PLS using the combined feature set was 0.75 whereas it was 0.72 for the baseline feature set.

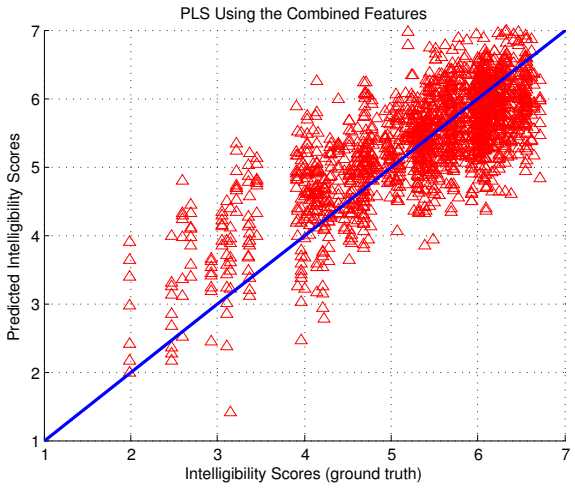


Figure 31: Ground-truth intelligibility scores against the predicted scores of PLS. The Pearson's r value is 0.75.

The ground-truth intelligibility scores and the predicted scores from PLS were quantized into 3 levels: poor, medium, and good, with the scores ranging between [1 3), [3 5), and [5 7], respectively. The predicted scores were obtained using the baseline, MRSTC, and the combined feature sets. The confusion matrices of the three regression models mapped into the three intelligibility levels are shown in Table 43. Each row represents the instances in an actual class normalized by the total number of the instances, and each column represents the normalized instances in a predicted class. Since the data are highly imbalanced (70% of the data values lie in the range between 5 and 7), unweighted accuracy (UWA) was chosen as the main evaluation measure as defined in Eq. (53).

The unweighted accuracies obtained by mapping the regression prediction scores were 59.5%, 53.2%, and 61.7% for the baseline, the MRSTC, and the combined feature sets, respectively. Although the unweighted accuracy of the MRSTC feature set is 6.3 percentage points below that of the baseline feature set, the unweighted accuracy is improved by

Table 43: Confusion matrices of the PLS regression predictions mapped into the 3 intelligibility levels. Chance would be 33.3.

Baseline				MRSTC				Baseline + MRSTC			
	poor'	med'	good'		poor'	med.'	good'		poor'	med'	good'
poor	26.8	73.2	0.0	poor	12.2	87.8	0.0	poor	29.3	70.7	0.0
med	0.9	61.4	37.7	med	1.6	56.4	42.0	med	1.3	65.1	33.6
good	0.0	10.4	89.6	good	0.0	12.8	87.2	good	0.0	9.5	90.5
UWA = 59.5%				UWA = 53.2%				UWA = 61.7%			

2.2 percentage points when those two feature sets were combined. The p-value for the improvement is less than 0.025 which indicates that the improvement is statistically significant.

A.3 Conclusion

The spectral features were evaluated for speech intelligibility prediction using the NKI CCRT Speech Corpus (NCSC). Head and neck cancer patients have communication difficulties due to hoarseness, dryness, excessive phlegm, coughing, and other articulation problems. Spectral analysis shows that hoarse voices consist of less conspicuous harmonics than normal voices. Other irregular and abnormal speech characteristics can be also observed in spectral analysis. Inclusion of these new features gave a 10% decrease in the mean squared error when tested using a partial least squares (PLS) regression algorithm with 10-fold cross validation. With a paired t-test, the decrease in the MSE was shown to be statistically significant with p-value less than 0.001. For the binary classification on the test set, a 5.0 percentage point improvement was observed when the new features were combined with the baseline features.

The typical STFT-based algorithms, which use a fixed analysis window, may fail to characterize harmonic characteristics in a broad frequency band. The multi-resolution approach enables better representation of spectral and harmonic characteristics by using longer windows in lower-frequency bands, and shorter windows in higher-frequency bands. By combining the new features with the baseline features, significant improvements in both MSE and unweighted accuracy were observed.

REFERENCES

[1] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.

[2] D. V. Anderson, "Speech analysis and coding using a multi-resolution sinusoidal transform," in *Acoustics, Speech, and Signal Processing, ICASSP. IEEE International Conference on, Atlanta, GA, USA*, vol. 2, pp. 1037–1040, IEEE, 1996.

[3] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[5] R. Calvo and S. D-Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[7] J. C. Kim, H. Rao, and M. A. Clements, "Investigating the use of formant based features for detection of affective dimensions in speech," in *Affective Computing and Intelligent Interaction, Memphis, TN, USA*, pp. 369–377, Springer, 2011.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[9] K. H. Kim, S. Bang, and S. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and Biological Engineering and Computing*, vol. 42, no. 3, pp. 419–427, 2004.

[10] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[11] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," *Semantic Computing, 2008 IEEE International Conference on, Santa Clara, CA, USA*, pp. 158–165, 2008.

[12] L. Vidrascu and L. Devillers, "Real-life emotion representation and detection in call centers data," *Affective Computing and Intelligent Interaction, Beijing, China*, vol. 3784, pp. 739–746, 2005.

[13] R. Sun, E. Moore, and J. F. Torres, "Investigating glottal parameters for differentiating emotional categories with similar prosodics," *Acoustics, Speech and Signal Processing, ICASSP 2009. IEEE International Conference on, Taipei, Taiwan*, pp. 4509–4512, 2009.

[14] Z.-J. Chuang and C.-H. Wu, "Multi-modal emotion recognition from speech and text," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, pp. 1–18, 2004.

[15] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on Hidden Markov Models," *Affective Computing and Intelligent Interaction, Memphis, TN, USA*, pp. 378–387, 2011.

[16] R. W. Picard, "Emotion research by the people, for the people," *Emotion Review*, vol. 2, no. 3, pp. 250–254, 2010.

[17] K. Takahashi and A. Tsukaguchi, "Remarks on emotion recognition from multi-modal bio-potential signals," *Systems, Man and Cybernetics, 2003. IEEE International Conference on, Washington, D.C., USA*, vol. 2, pp. 1654–1659, 2003.

[18] C. Darwin, *The Expression of the Emotions in Man and Animals*. D. Appleton, 1898.

[19] W. James, "What is an emotion?," *Mind*, no. 34, pp. 188–205, 1884.

[20] M. B. Arnold, *Emotion and Personality*. Columbia University Press, 1960.

[21] J. R. Averill, "A constructivist view of emotion," *Emotion: Theory, Research, and Experience*, vol. 1, pp. 305–339, 1980.

[22] P. M. Cole, C. J. Bruschi, and B. L. Tamang, "Cultural differences in children's emotional reactions to difficult situations," *Child Development*, vol. 73, no. 3, pp. 983–996, 2002.

[23] J. J. Gross and R. A. Thompson, "Emotion regulation: Conceptual foundations," *Handbook of Emotion Regulation*, vol. 3, p. 24, 2007.

[24] J. LeDoux, *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon and Schuster, 1998.

[25] J. A. Russell, "Core affect and the psychological construction of emotion.," *Psychological Review*, vol. 110, no. 1, pp. 145–172, 2003.

[26] T. Bänziger and K. R. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus," in *Affective Computing and Intelligent Interaction, Lisbon, Portugal*, pp. 476–487, Springer, 2007.

[27] T. Bänziger, H. Pirker, and K. Scherer, "GEMEP-Geneva multimodal emotion portrayals: A corpus for the study of multimodal emotional expressions," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC), Genoa, Italy*, vol. 6, pp. 15–019, 2006.

[28] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[29] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[30] L. Devillers and I. Vasilescu, "Reliability of lexical and prosodic cues in two real-life spoken dialog corpora," *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal*, 2004.

[31] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *Image and Vision Computing*, vol. 25, no. 12, pp. 1856–1863, 2007.

[32] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands*, pp. 317–321, 2005.

[33] A. J. O'Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi, "A video database of moving faces and people," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 812–816, 2005.

[34] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. DArcy, M. Russell, and M. Wong, "you stupid tin box-children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," *Proceedings of the International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal*, pp. 171–174, 2004.

[35] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, *et al.*, "Distinguishing deceptive from non-deceptive speech," *Proc. Eurospeech, Lisbon, Portugal*, pp. 1833–1836, 2005.

[36] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA*, vol. 2, pp. 568–573, 2005.

[37] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," *IEEE International Conference on Multimedia and Expo, Singapore*, pp. 1079–1084, 2010.

[38] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1, pp. 33–60, 2003.

[39] G. I. Roisman, J. L. Tsai, K.-H. S. Chiang, *et al.*, "The emotional integration of childhood experience: Physiological, facial expressive, and self-reported emotional response during the adult attachment interview," *Developmental Psychology*, vol. 40, no. 5, pp. 776–789, 2004.

[40] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," *7th International Conference on Spoken Language Processing (ICSLP), Denver, Colorado, USA*, vol. 2, pp. 301–304, 2002.

[41] I. S. Engberg and A. V. Hansen, "Documentation of the Danish emotional speech database DES," *Internal AAU Report, Center for Person Kommunikation, Denmark*, 1996.

[42] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using Hidden Markov Models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, 2003.

[43] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," *Proceedings of the International Conference on Multimedia, Singapore*, pp. 1459–1462, 2010.

[44] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," *Proc. Interspeech 2013, ISCA, Lyon, France*, 2013.

[45] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, *et al.*, "The Interspeech 2012 speaker trait challenge," *Proc. Interspeech 2012, ISCA, Portland, OR, USA*, 2012.

[46] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[47] X. Wang and K. K. Paliwal, "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern Recognition*, vol. 36, no. 10, pp. 2429–2439, 2003.

[48] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features," *IEEE 9th Workshop on Multimedia Signal Processing, Chania, Crete, Greece*, pp. 48–51, 2007.

[49] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," *Proc. Eurospeech, Lisbon, Portugal*, pp. 497–500, 2005.

[50] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Interspeech 2009, Brighton, UK*, pp. 320–323, 2009.

[51] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," *Interspeech 2009, Brighton, UK*, pp. 312–315, 2009.

[52] C.-H. Wu and W.-B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2011.

[53] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 116–125, 2012.

[54] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *Proc. Interspeech 2005, Lisbon, Portugal*, 2005.

[55] A. Batliner, S. Steidl, F. Eyben, and B. Schuller, "On laughter and speech laugh, based on observations of child-robot interaction," *The Phonetics of Laughing*, 2011.

[56] P. R. Kleinginna Jr and A. M. Kleinginna, "A categorized list of emotion definitions, with suggestions for a consensual definition," *Motivation and Emotion*, vol. 5, no. 4, pp. 345–379, 1981.

[57] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[58] D. France, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[59] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," *Engineering in Medicine and Biology Society, 2004. IEMBS '04. 26th Annual International Conference of the IEEE, San Francisco, CA, USA*, vol. 1, pp. 17–20, 2004.

[60] W. R. Rodrguez and E. Lleida, "Formant estimation in childrens speech and its application for a Spanish speech therapy tool," *Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE), Warwickshire, UK*, 2009.

[61] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Perception and Psychophysics*, vol. 27, no. 1, pp. 24–27, 1980.

[62] R. W. Frick, "The prosodic expression of anger: Differentiating threat and frustration," *Aggressive Behavior*, vol. 12, no. 2, pp. 121–128, 1986.

[63] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[64] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction, Memphis, TN, USA*, pp. 415–424, Springer, 2011.

[65] A. Mehrabian, *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Oelgeschlager, Gunn & Hain Cambridge, MA, 1980.

[66] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[67] A. V. Oppenheim, R. W. Schafer, J. R. Buck, *et al.*, *Discrete-Time Signal Processing*, vol. 2. Prentice-hall Englewood Cliffs, 1989.

[68] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[69] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 02, pp. 185–205, 2005.

[70] R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz, "Best agglomerative ranked subset for feature selection.," in *Workshop on New Challenge for Feature Selection in Data Mining and Knowledge Discovery (FSDM), Antwerp, Belgium*, pp. 148–162, 2008.

[71] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, pp. 3795–3806, 1996.

[72] K. K. Paliwal and W. Kleijn, "Quantization of LPC parameters," *Speech Coding and Synthesis*, pp. 433–466, 1995.

[73] J. N. Holmes, W. J. Holmes, and P. N. Garner, "Using formant frequencies in speech recognition," *Proc. European Conf. on Speech Communication and Technology, Rhodes, Greece*, vol. 4, pp. 2083–2086, 1997.

[74] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of Gaussians," *Acoustics, Speech, and Signal Processing, ICASSP. IEEE International Conference on, Atlanta, GA, USA*, vol. 2, pp. 1229–1232, 1996.

[75] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[76] J. C. Kim, H. Rao, and M. A. Clements, "Formant frequency tracking using Gaussian mixtures with maximum a posteriori adaptation," in *Proc. Interspeech 2013, ISCA, Lyon, France*, pp. 3221–3225, 2013.

[77] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[78] M. Stuttle, "A Gaussian mixture model spectral representation for speech recognition," *Ph.D. Thesis, Engineering Department, University of Cambridge*, 2003.

[79] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[80] J. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[81] L. Deng, X. Cui, P. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," *Acoustics, Speech, and Signal Processing, ICASSP. IEEE International Conference on, Toulouse, France*, vol. 1, pp. 60–63, 2006.

[82] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. on Speech and Audio Processing*, pp. 435–444, Mar. 2006.

[83] J. Tierney, "A study of LPC analysis of speech in additive noise," *IEEE Trans on Speech and Audio Processing*, pp. 389–397, Aug. 1980.

[84] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: A stepwise procedure for building and training a neural network," in *Neurocomputing*, pp. 41–50, Springer, 1990.

[85] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[86] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7, pp. 613–625, 2010.

[87] T. F. Quatieri and R. McAulay, "Shape invariant time-scale and pitch modification of speech," *Signal Processing, IEEE Transactions on*, vol. 40, no. 3, pp. 497–510, 1992.

[88] J. C. Kim and M. A. Clements, "Time-scale modification of audio signals using multi-relative onset time estimations in sinusoidal transform coding," *Signals, Systems and Computers (ASILOMAR)*, pp. 558–561, 2010.

[89] D. B. Paul, "The spectral envelope estimation vocoder," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 4, pp. 786–794, 1981.

[90] T. Quatieri, *Discrete-Time Speech Signal Processing*. Prentice Hall, 2002.

[91] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909–996, 1988.

[92] D. V. Anderson and M. A. Clements, "Efficient multi-resolution sinusoidal modeling," *World Multiconference on Systemics, Cybernetics, and Informatics, Orlando, FL, USA*, vol. 6, pp. 424–429, 2000.

[93] H. Hermansky and N. Morgan, "RASTA processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, 1994.

[94] S. G. Koolagudi and K. S. Rao, "Exploring speech features for classifying emotions along valence dimension," in *Pattern Recognition and Machine Intelligence*, pp. 537–542, Springer, 2009.

[95] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," *Proceedings of the 6th International Conference on Multimodal Interfaces, State College, PA, USA*, pp. 205–211, 2004.

[96] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *The Quarterly Journal of Experimental Psychology*, vol. 63, no. 11, pp. 2251–2272, 2010.

[97] E. Lasarcyk and J. Trouvain, "Spread lips + raised larynx + higher F0 = Smiled Speech?-An articulatory synthesis approach," *8th International Seminar on Speech Production, Strasbourg, France*, pp. 345–348, 2008.

[98] X. Li, J. Tao, M. Johnson, J. Soltis, A. Savage, K. Leong, and J. Newman, "Stress and emotion classification using jitter and shimmer features," *IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawaii, USA*, vol. 4, pp. IV–1081–IV–1084, 2007.

[99] S.-S. Choi, S.-H. Cha, and C. C. Tappert, "A survey of binary similarity and distance measures," *Journal of Systemics, Cybernetics and Informatics*, vol. 8, no. 1, pp. 43–48, 2010.

[100] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[101] T. Li, "A general model for clustering binary data," *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, IL, USA*, pp. 188–197, 2005.

[102] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 568–586, 2011.

[103] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[104] J. R. Williams, "Guidelines for the use of multimedia in instruction," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Chicago, IL, USA*, vol. 42, no. 20, pp. 1447–1451, 1998.

[105] R. Sun and E. I. Moore, "Empirical study of dimensional and categorical emotion descriptors in emotional speech perception," *Florida Artificial Intelligence Research Society Conference, Marco Island, Florida, USA*, pp. 104–109, 2012.

[106] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, Vancouver, Canada*, pp. 3687–3691, IEEE, 2013.

[107] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, pp. 1–23, 2012.

[108] O. AlZoubi, S. K. D'Mello, and R. A. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *Affective Computing, IEEE Transactions on*, vol. 3, no. 3, pp. 298–310, 2012.

[109] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[110] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[111] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in *Data Mining Techniques for the Life Sciences*, pp. 223–239, Springer, 2010.

[112] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[113] B. Schuller, M. Wöllmer, F. Eyben, and G. Rigoll, "Spectral or voice quality? feature type relevance for the discrimination of emotion pairs," *The Role of Prosody in Affective Speech. Linguistic Insights, Studies in Language and Communication*, vol. 97, pp. 285–307, 2009.

[114] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[115] S. Fagel, "Effects of smiled speech on lips, larynx and acoustics," *Lecture Notes in Computer Science*, vol. 5967, pp. 294–303, 2010.

[116] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of Nonverbal Behavior*, vol. 6, no. 4, pp. 238–252, 1982.

[117] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1, pp. 227–256, 2003.

[118] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *Journal of Personality and Social Psychology*, vol. 70, no. 3, p. 614, 1996.

[119] K. Tamuri, "Intensity of estonian emotional speech.," in *Baltic Human Language Technologies 2012, Tartu, Estonia*, pp. 238–246, 2012.

[120] D. P. Szameitat, K. Alter, A. J. Szameitat, D. Wildgruber, A. Sterr, and C. J. Darwin, "Acoustic profiles of distinct emotional expressions in laughter," *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 354–366, 2009.

[121] T. Johnstone and K. R. Scherer, "The effects of emotions on voice quality," in *Proceedings of the XIVth International Congress of Phonetic Sciences*, pp. 2029–2032, University of California, Berkeley, San Francisco, CA, 1999.

[122] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[123] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernández-Dols, "Facial and vocal expressions of emotion," *Annual Review of Psychology*, vol. 54, no. 1, pp. 329–349, 2003.

[124] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[125] K. Goebel, W. Yan, and W. Cheetham, "A method to calculate classifier correlation for decision fusion," in *Proceedings of Information, Decision, and Control, Las Vegas, NV*, vol. 2002, pp. 135–140, 2002.

[126] A. M. Kring and A. H. Gordon, "Sex differences in emotion: expression, experience, and physiology.," *Journal of Personality and Social Psychology*, vol. 74, no. 3, p. 686, 1998.

[127] R. W. Simon and L. E. Nath, "Gender and emotion in the united states: Do men and women differ in self-reports of feelings and expressive behavior?," *American Journal of Sociology*, vol. 109, no. 5, pp. 1137–1176, 2004.

[128] H. Rao, J. C. Kim, A. Rozga, and M. A. Clements, "Detection of laughter in children's speech using spectral and prosodic acoustic features," in *Proc. Interspeech 2013, ISCA, Lyon, France*, pp. 1399–1403, 2013.

[129] R. Gupta, C.-C. Lee, and S. Narayanan, "Classification of emotional content of sighs in dyadic human interactions," *IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan*, pp. 2265–2268, 2012.

[130] M. J. Owren and J.-A. Bachorowski, "Reconsidering the evolution of nonlinguistic communication: The case of laughter," *Journal of Nonverbal Behavior*, vol. 27, no. 3, pp. 183–200, 2003.

[131] A. P. Nilsen and D. L. Nilsen, *Encyclopedia of 20th-Century American Humor.* ERIC, 2000.

[132] K. H. Teigen, "Is a sigh 'just a sigh'? Sighs as emotional signals and responses to a difficult task," *Scandinavian Journal of Psychology*, vol. 49, no. 1, pp. 49–57, 2008.

[133] S. J. Webb, E. J. Jones, J. Kelly, and G. Dawson, "The motivation for very early intervention for infants at high risk for autism spectrum disorders," *International Journal of Speech-Language Pathology*, vol. 16, no. 1, pp. 36–42, 2014.

[134] J. M. Rehg, A. Rozga, G. D. Abowd, and M. S. Goodwin, "Behavioral imaging and autism," *Pervasive Computing, IEEE*, vol. 13, no. 2, pp. 84–87, 2014.

[135] L. K. Koegel, A. K. Singh, R. L. Koegel, J. R. Hollingsworth, and J. Bradshaw, "Assessing and improving early social engagement in infants," *Journal of Positive Behavior Interventions*, vol. 16, no. 2, pp. 69–80, 2014.

[136] L. Zwaigenbaum, S. Bryson, and N. Garon, "Early identification of autism spectrum disorders," *Behavioural Brain Research*, vol. 251, pp. 133–146, 2013.

[137] B. A. Corbett, C. Newsom, A. P. Key, L. R. Qualls, and E. K. Edmiston, "Examining the relationship between face processing and social interaction behavior in children with and without autism spectrum disorder," *Journal of Neurodevelopmental Disorders*, vol. 6, no. 1, pp. 1–11, 2014.

[138] R. Gupta, C.-C. Lee, D. Bone, A. Rozga, S. Lee, and S. Narayanan, "Acoustical analysis of engagement behavior in children," in *Third Workshop on Child, Computer and Interaction, Portland, OR, USA*, 2012.

[139] J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, and R. W. Picard, "Using electrodermal activity to recognize ease of engagement in children during social interactions," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA*, pp. 307–317, ACM, 2014.

[140] J. Whitehill, Z. Serpell, A. Foster, Y.-C. Lin, B. Pearson, M. Bartlett, and J. Movellan, "Towards an optimal affect-sensitive instructional system of cognitive skills," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on, Colorado Springs, CO*, pp. 20–25, IEEE, 2011.

[141] J. C. Kim and M. A. Clements, "Multimodal affect classification at various temporal lengths," *Affective Computing, IEEE Transactions on*, 2014, (under revision).

[142] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, H. Rao, J. Kim, L. Presti, J. Zhang, L. D., B. J., and Z. Ye, "Decoding children's social behavior," *Computer Vision and Pattern Recognition (CVPR), Portland, OR*, 2013.

[143] J. C. Kim, H. Rao, A. Rozga, and M. A. Clements, "Social engagement classification in dyadic plays," *Affective Computing, IEEE Transactions on*, 2014, (under review).

[144] S. J. Sheinkopf, P. Mundy, A. H. Claussen, and J. Willoughby, "Infant joint attention skill and preschool behavioral outcomes in at-risk children," *Development and Psychopathology*, vol. 16, no. 02, pp. 273–291, 2004.

[145] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC), Genoa, Italy*, vol. 2006, 2006.

[146] R. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on, Albuquerque, NM, USA*, pp. 249–252, IEEE, 1990.

[147] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[148] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, no. 01, pp. 1–63, 2000.

[149] H. Sacks, *Lectures on Conversation*, vol. 1. Blackwell Publishing, 1995.

[150] H. Rao, J. C. Kim, M. A. Clements, A. Rozga, and D. S. Messinger, "Detection of children's paralinguistic events in interaction with caregivers," in *Proc. Interspeech 2014, ISCA, Singapore*, 2014.

[151] H. Goodluck and X. Li, *Language Acquisition: A Linguistic Introduction*. Blackwell Cambridge, MA, 1991.

[152] R. C. Anderson and K. A. Franke, "Psychological and psychosocial implications of head and neck cancer.," *Internet Journal of Mental Health*, vol. 1, no. 2, 2002.

[153] M. Björklund, A. Sarvimäki, and A. Berg, "Health promotion and empowerment from the perspective of individuals living with head and neck cancer," *European Journal of Oncology Nursing*, vol. 12, no. 1, pp. 26–34, 2008.

[154] D. M. Brizel, T. H. Wasserman, M. Henke, V. Strnad, V. Rudat, A. Monnier, F. Eschwege, J. Zhang, L. Russell, W. Oster, *et al.*, "Phase III randomized trial of amifostine as a radioprotector in head and neck cancer," *Journal of Clinical Oncology*, vol. 18, no. 19, pp. 3339–3345, 2000.

[155] H.-J. Guchelaar, A. Vermes, and J. H. Meerwaldt, "Radiation-induced xerostomia: pathophysiology, clinical course and supportive treatment," *Supportive Care in Cancer*, vol. 5, no. 4, pp. 281–288, 1997.

[156] L. Gavidia-Ceballos and J. H. L. Hansen, "Direct speech feature estimation using an iterative EM algorithm for vocal fold pathology detection," *Biomedical Engineering, IEEE Transactions on*, vol. 43, no. 4, pp. 373–383, 1996.

[157] L. Van Der Molen, M. van Rossum, A. Ackerstaff, L. Smeele, C. Rasch, and F. Hilgers, "Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views," *BMC Ear, Nose and Throat Disorders*, vol. 9, no. 1, p. 10, 2009.

[158] R. P. Clapham, L. van der Molen, R. van Son, M. van den Brekel, and F. J. Hilgers, "NKI-CCRT corpus: speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy," *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey*, 2012.